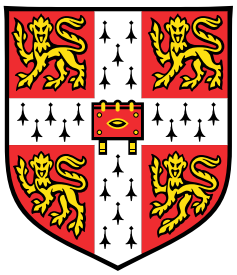
Modelling longitudinal data on respiratory infections to inform health policy



Chiara Chiavenna

MRC Biostatistics Unit
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Emmanuel College June 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Chiara Chiavenna
June 2020

Abstract

Detecting the start of an outbreak, quantifying its burden, disentangling the contribution of different pathogens and evaluating the effectiveness of an intervention are research questions common to several infectious diseases. The answers to these questions provide the epidemiological understanding to prevent future outbreaks, by informing public health policies such as drug stockpiling, vaccination regimes or non-medical interventions. We investigate the use of statistical models to quantify burden of respiratory disease and evaluate effectiveness of public health interventions, while accounting for the challenges posed by surveillance data. The observational nature of the available information, affected by confounding, makes causal statements difficult. Improvements to routinely employed methodologies are proposed, employing phenomenological models to estimate a counterfactual, i.e. what what would have happened in the absence of a contributing factor or intervention.

We apply these methods to different types of studies, to address specific gaps in the literature. *S. pneumoniae* is the leading cause of respiratory morbidity and mortality globally, especially in young children and in the elderly. To improve the understanding of factors triggering disease progression, we firstly analyse individual-level information about pneumococcal carriage and lower respiratory tract infection with a multi-state model, using data from a cohort study in Thailand. Secondly, we clarify the role of viral coinfection and meteorological conditions in invasive pneumococcal disease (IPD) incidence using English surveillance data. A novel multivariate linear regression model is proposed to estimate the influenza-specific contribution additional to the seasonal IPD burden across age groups. We then quantify the impact of the currently implemented vaccination policy, by estimating the counterfactual of IPD incidence in absence of vaccination. This allows disentangling serotype replacement from the vaccine effect, making use of a synthetic control approach. Finally, an empirical dynamical modelling strategy is employed to quantify the interaction between influenza and pneumococcus. Counterfactual analysis can also be employed to quantify the burden of novel respiratory pathogens. The last application of this approach is to estimate the excess mortality during the the COVID-19 pandemic in England.

Acknowledgements

I would like to express my gratitude to my supervisors, Prof. Daniela De Angelis and Dr Anne Presanis. Thank you Dani for giving me the opportunity to undertake this project, for granting me the independence to explore my own research interests, for your attention to detail and for the many stimulating questions that encouraged me to improve this research over the past four years. Thank you Anne for your patient guidance, constant encouragement, precious advice and thoughtful revisions.

I am grateful to Public Health England for the financial support, and to Dr Richard Pebody and Dr Andre Charlett for providing essential data and useful discussion, along with Professor Simon de Lusignan. Daniela, Anne, Richard, Andre and Simon are co-authors on a paper based on the work in Chapter 4, however the statistical analysis and text presented in this thesis are my own. Thanks also to Prof. Paul and Claudia Turner for sharing the data presented in chapter 2, and to all the healthcare workers who collected these data. Finally, thanks to the friendly scientists and helpful computing staff at the Biostatistics Unit for the excellent statistical discussions and for always being willing to answer my questions.

Special thanks go to the energising friends who made my life in Cambridge a lot more pleasant. I would like to acknowledge in particular Maria, Delphine, Michela, Gian, Silvia, Georgia, Aida, Lucia, Ennio, Emily, Abdullah and Ana. You have been the best local family I could ask for, and this thesis would have never been finished without your support. Thanks also to Paolo for the mentoring and proof-reading, to my friends around the world Alice, Lan, Claire, Nico, Ottavia, Sissi, Alex, Erika, Cris and Stefi, and last but not least to my family. Every message and every call of encouragement meant a lot.

Finally, thanks to my examiners for the enjoyable discussion during the viva and for the positive and constructive feedback on this thesis.

Table of contents

List of figures **xiii**

List of tables **xvii**

1 Introduction **1**
1.1 History of epidemics 1
1.2 Epidemiology and public health of outbreaks 3
1.3 Statistical models for epidemic data 5
1.4 Respiratory infections and co-infections 6
1.5 Challenges in estimating LRTI burden 7
1.6 Aims of the thesis 10

2 Individual-level dynamics of disease **13**
2.1 Introduction 13
2.2 Multistate models 14
2.2.1 Notation and assumptions 14
2.2.2 Estimating transitions using panel data 15
2.2.3 Mixture of observed and unobserved transition times 16
2.3 Application to pneumococcal disease 17
2.3.1 Previous multistate models for S. pneumoniae 17
2.3.2 Source of data 18
2.3.3 Analysis strategy 18
2.4 Results 20
2.4.1 Descriptive analysis 20
2.4.2 Multistate model 22
2.5 Discussion 27

3 Models for burden estimation from time series counts **31**
3.1 Introduction 31

3.2 Ecological studies . . . 31
3.2.1 Cyclic regression models . . . 33
3.2.2 Time series methodology . . . 35
3.2.3 Poisson branching processes . . . 38
3.3 Predictive model assessment . . . 40
3.4 Granger causality . . . 43
3.5 Conclusions . . . 44

4 Estimating age-stratified influenza-associated IPD in England 45
4.1 Introduction . . . 45
4.2 Data . . . 46
4.2.1 Bivariate time series analysis . . . 49
4.3 Analysis strategy . . . 51
4.4 Results . . . 53
4.4.1 Model choice: endemic waves and lagged covariates . . . 53
4.4.2 Estimated influenza impact . . . 54
4.4.3 Rhinovirus and RSV . . . 54
4.4.4 Age-specific analysis . . . 55
4.5 Discussion . . . 61

5 Time-series methods to assess the impact of an intervention 65
5.1 Introduction . . . 65
5.2 Intervention evaluation framework . . . 66
5.3 Before-and-after designs . . . 66
5.3.1 Interrupted time series analysis . . . 67
5.4 Controlled designs . . . 68
5.4.1 Synthetic controls . . . 69
5.4.2 Controlled interrupted time series (CITS) . . . 70
5.4.3 The Causal Impact method (CIM) . . . 71
5.5 Predictive model assessment . . . 75
5.6 Conclusions . . . 76

6 Application to Pneumococcal Conjugate Vaccine 79
6.1 Introduction . . . 79
6.2 Data . . . 81
6.2.1 Selection of controls . . . 85
6.3 Analysis strategy . . . 87

6.3.1 ITS regression . . . 87
6.3.2 CIM . . . 88
6.3.3 ITS analysis . . . 89
6.3.4 CIM . . . 93
6.4 Discussion . . . 99

7 Estimated excess all-cause mortality in England during the COVID-19 pandemic 107
7.1 Introduction . . . 107
7.2 Data . . . 109
7.2.1 All-cause deaths . . . 109
7.2.2 COVID-lab-confirmed deaths . . . 111
7.3 Analysis strategy . . . 111
7.4 Results . . . 114
7.4.1 Results by gender . . . 115
7.4.2 Results by age . . . 116
7.4.3 Results by region . . . 118
7.4.4 Comparison with Poisson regression . . . 118
7.5 Discussion . . . 123

8 Empirical dynamical modelling 129
8.1 Introduction . . . 129
8.1.1 Why a nonlinear time series analysis? . . . 130
8.1.2 EDM in infectious disease . . . 132
8.2 Methods . . . 133
8.2.1 Dynamical systems and their geometry . . . 133
8.2.2 Phase space reconstruction techniques . . . 135
8.2.3 Convergent Cross Mapping (CCM) . . . 137
8.2.4 Signal processing techniques . . . 142
8.3 Application to influenza and IPD time series . . . 144
8.3.1 SSA for IPD . . . 144
8.3.2 SSA for influenza . . . 147
8.3.3 Embedding of IPD and influenza signals . . . 149
8.3.4 CCM of influenza and IPD . . . 151
8.4 Discussion . . . 153

9 Conclusions and future work **155**
9.1 Main thesis findings . . . 156
9.1.1 Pneumococcal disease progression . . . 156
9.1.2 Burden of endemic respiratory pathogens . . . 156
9.1.3 Evaluation of pneumococcal vaccine and serotype replacement . . . 157
9.1.4 COVID-19 excess mortality . . . 158
9.1.5 Empirical dynamical modelling . . . 159
9.2 Future work . . . 159
9.2.1 Extensions to the hhh modelling framework . . . 159
9.2.2 Incorporating a transmission model . . . 160
9.2.3 Multiple baseline design . . . 160
9.2.4 Indirect effect of reduced S.Pneumoniae circulation . . . 161
9.3 Concluding remarks . . . 161

References **163**

Appendix A Supplementary information to chapter 4 **181**

Appendix B Supplementary information to chapter 6 **191**
B.1 Observed IPD incidence by age and PCV group . . . 191
B.2 Sensitivity analysis for change of lag in ITS . . . 195
B.3 BSTS for IPD by age . . . 199
B.3.1 Models B and C: impact of PCV7 . . . 199
B.3.2 Models D and E: impact of PCV13 . . . 200

Appendix C Supplementary information to chapter 7 **211**
C.1 Singular value decomposition (SVD) . . . 211
C.2 Supplementary results . . . 212
C.2.1 SSA IPD . . . 212
C.2.2 SSA flu . . . 213
C.2.3 Embedding and CCM on original rates . . . 213

List of figures

1.1 Process of bacterial progression from upper to lower respiratory tract triggered by viral infection [132] 9

2.1 Example of process observed under a panel data scheme 16
2.2 Transitions between states that are instantaneously possible 19
2.3 Weather conditions, carriage prevalence, median cohort age, pneumonia incidence and viral circulation. 21
2.4 Viral positivity obtained from the national surveillance system of Thailand. 22
2.5 In the first three panels: observed and estimated prevalence across different states over study time. In the last panel: survival function for time to pneumonia. 24

4.1 ILI and Flu incidence rate in the top panel, IPD incidence rate in the bottom one. 48
4.2 Cross-correlation between each virus and IPD for up to 15 weeks of lag. Top panel: flu and IPD. Middle panel: RSV and IPD. Bottom panel: rhinovirus and IPD. 49
4.3 Cross-correlation between weather variables and IPD for up to 15 weeks of lag. 50
4.4 Model (B) of IPD and Influenza with one set of trigonometric functions . . 55
4.5 Model (D) of IPD and Influenza with rainfall and temperature 55
4.6 One-step-ahead predictive distribution and observed values 56
4.7 Model (F) including Influenza, rhinovirus and RSV 57
4.8 Model I: Fitted IPD values for infants and school-age children 58
4.9 Model I: Fitted IPD values for the 15-44 and 45-64 age groups 59
4.10 Model I: Fitted IPD values for the elderly 60

5.1 Difference-in-difference estimation, graphical explanation [Columbia University] 68
5.2 Graphical representation of evaluation on a rolling forecasting origin 76

6.1 Monthly IPD incidence rate per million residents 82

6.2 Monthly age-specific IPD incidence rate per million residents 83
6.3 Monthly IPD incidence rate per million residents by serotype 84
6.4 Control time series - incidence rate per million residents 86
6.5 Model A: fitted IPD counts based on three ITS models 89
6.6 Top three panels: model B, fitted PCV7-IPD counts. Bottom three panels: model C, fitted non-PCV7-IPD counts, based on three ITS models 91
6.7 Top three panels: model D, fitted IPD-PCV13 counts. Bottom three panels: model E, fitted IPD-NVT counts, based on three ITS models 92
6.8 Heatmap of MSPE for different state models and training sets 93
6.9 Model A: impact of PCV introduction on the overall IPD incidence rate . . 94
6.10 Posterior probability of inclusion for each control time series 95
6.11 Model A: impact of PCV introduction on the overall IPD incidence rate in children younger than 5 (top panel) and aged 5-14 (bottom panel). 97
6.12 Model A: impact of PCV introduction on the overall IPD incidence rate in adults aged 15-44 (top panel) and 45-64 (bottom panel). 98
6.13 Model A: impact of PCV introduction on the overall IPD incidence rate in the elderly (65+). 99
6.14 Top panel: model B, impact of PCV introduction on the PCV7-IPD incidence rate. Bottom panel: model C, impact of PCV7 introduction on the nonPCV7-IPD incidence rates. 100
6.15 Top panel: model D, impact of PCV13 introduction on the PCV13-IPD incidence rate. Bottom panel: model E, impact of PCV13 introduction on the NVT-IPD incidence rates. 102

7.1 Observed mortality rate in the past 5 epidemic years in England 110
7.2 Observed mortality rate in the past 5 epidemic years in England by gender . 111
7.3 Observed mortality rate in the past 5 epidemic years in England by region . 112
7.4 Observed mortality rate in the past 5 epidemic years in England by age group 113
7.5 Posterior probability of inclusion for each control time series 115
7.6 Excess all-cause mortality for men (top panel) and women (bottom one) . . 117
7.7 Daily excess in all-cause mortality by age group, per 100'000 residents . . 119
7.8 Excess all-cause mortality in England (top panel) and in London only (bottom panel) 120
7.9 Excess all-cause mortality for North West, North East, East Midlands and East of England 121
7.10 Excess all-cause mortality for West Midlands, Yorkshire and the Humber, South West and South East 122

7.11 Observed all-cause deaths and estimated counterfactual obtained with the BSTS model (top panel) and with the Poisson regression (bottom panel) for England up to May 29th 124

8.1 Correlations between all pairs of variables for the Lorenz attractor [198] . . 131
8.2 Top: representation of a Lorenz system trajectory in the x, y, z space, where the variables are temperature, pressure gradient and angular velocity. Middle: time series of $x(t)$, discarding any knowledge of y, z, or the governing equations. Bottom: phase-space reconstructed by embedding $x(t)$. 134
8.3 Graphical representation of the correspondence of points on manifolds M, M_x and M_y as presented by Sugihara et al. [199] 138
8.4 Effect of time delays on cross mapping. Panel (A) shows causation for two cases: (i) no time delay in the effect of x on y and (ii) y responds to x with a time delay of 4 time steps. Panel (B) shows (i) cross mapping with l=0, equivalent to the original formulation by [198] and (ii) cross mapping with l=-4, which may be expected to be better than l=0 when x acts on y with some time delay. [237] 141
8.5 Eigenvalue spectrum and W-correlation matrix for IPD 144
8.6 Reconstructed elementary components for IPD 145
8.7 Signal for IPD reconstructed with SSA 146
8.8 Eigenvalue spectrum and W-correlation matrix for influenza 147
8.9 Reconstructed elementary components for flu 148
8.10 Signal for influenza reconstructed with SSA 149
8.11 Selection of time delay τ and embedding dimension E for the IPD signal . . 150
8.12 Selection of time delay τ and embedding dimension E for the flu signal . . 150
8.13 Reconstructed manifolds for IPD signal (left) and influenza signal (right) . . 151
8.14 Correlation coefficients for prediction skills of time-delayed CCM 152

A.1 Fitted IPD values all ages 181
A.2 RSV and rhinovirus incidence rates 182
A.3 Observed counts: 15-44 years old 183
A.4 Observed counts: 45-64 years old 184
A.5 Observed counts: 65+ years old 185
A.6 Model H: Predictive distribution for infants and young adults 186
A.7 Model H: Predictive distribution for 45-64 and 65+ 187
A.8 Model K: Fitted IPD values for infants 188
A.9 Model K: Fitted IPD values for school-age children 188

A.10 Model K: Fitted IPD values for young adults 189
A.11 Model K: Fitted IPD values for the 45-64 age group 189
A.12 Model K: Fitted IPD values for the elderly 190

B.1 PCV7 serotypes, incidence rate per million residents (scales differ across panels) 192
B.2 PCV13 serotypes, incidence rate per million residents (scales differ across panels) 193
B.3 NV serotypes, incidence rate per million residents (scales differ across panels) 194
B.4 Control time series, incidence rate per million residents (scales differ across panels) 195
B.5 Fitted IPD counts based on three ITS models 196
B.6 Fitted PCV7- and nonPCV7-IPD counts based on three ITS models 197
B.7 Fitted PCV13- and NVT-IPD counts based on three ITS models 198
B.8 Fitted PCV7- and nonPCV7-IPD incidence rates in children younger than 5 200
B.9 Fitted PCV7- and nonPCV7-IPD incidence rates in children of age 5-14 . . 201
B.10 Fitted PCV7- and nonPCV7-IPD incidence rates in adults aged 15-44 . . . 202
B.11 Fitted PCV7- and nonPCV7-IPD incidence rates in adults aged 45-64 . . . 203
B.12 Fitted PCV7- and nonPCV7-IPD incidence rates in adults aged 65+ 204
B.13 Fitted PCV13- and NVT-IPD incidence rates in children younger than 5 . . 205
B.14 Fitted PCV13- and NVT-IPD incidence rates in children of age 5-14 206
B.15 Fitted PCV13- and NVT-IPD incidence rates in adults aged 15-44 207
B.16 Fitted PCV13- and NVT-IPD incidence rates in adults aged 45-64 208
B.17 Fitted PCV13- and NVT-IPD incidence rates in adults aged 65+ 209

C.1 Eigenvectors for SSA of IPD time series, individually and in pairs 214
C.2 W-correlation matrix on the IPD residuals shows no separability for other components 215
C.3 Eigenvectors for SSA of Flu time series, individually and in pairs 216
C.4 W-correlation matrix on the IPD residuals shows no separability for other components 217
C.5 Selection of time delay τ and embedding dimension E for the original IPD rates 217
C.6 Selection of time delay τ and embedding dimension E for the original flu rates 218
C.7 CCM predictive skills when applied on the original time series 219

List of tables

2.1 Observed transitions across states from the dataset considered. 23
2.2 Estimated transition intensities (and 95% confidence intervals) for the multi-state model in Figure 2.2. 23
2.3 Estimated mean sojourn for each state visit, expected number of visits to each state in the two years of follow-up, and forecast total length of time spent in each state in the two years of follow-up for the multistate model in Figure 2.2. 24
2.4 Model selection steps: LR test for addition of covariates to the model in Figure 2.2. 25
2.5 Estimated HRs (95% CI) for the covariates included in the selected model . 26
2.6 Estimated transition intensities (and 95% confidence intervals) for dry season with no Influenza, vs wet season with Influenza 29

4.1 P-values from a Granger causality test assessing the predictability of IPD as a function the listed covariates 51
4.2 Model comparison in terms of AIC and one-step ahead forecast (log(s(P,x))) 53
4.3 Coefficient estimates for model (I), including Flu, rhinovirus and RSV as covariates. 56
4.4 Multivariate model comparison in terms of AIC 57
4.5 Model I: Coefficient estimates for the age-specific model of IPD including Flu. Since Flu coefficients $\tau_{<5}$ and τ_{65+} were very small, we refit the model fixing them to 0, to make sure the other parameter estimates are not sensitive to such an assumption. 60
4.6 Model I: Standard error estimates for the age-specific model of IPD including Flu. Uncertainty around coefficients $\tau_{<5}$ and τ_{65+} was not well estimated. . 61
4.7 Model I: Relative proportions (%) of IPD cases attributed to pneumococcal transmission within and across age groups, and to influenza overall or in the pandemic period 61

4.8 Model I: Coefficient estimates for the age-specific model of IPD including Flu 62
4.9 Model I: Coefficient standard errors for the age-specific model of IPD including Flu 62
4.10 Model K: Relative proportions (%) of IPD cases attributed to pneumococcal transmission within and across age groups, to influenza, rhinovirus and RSV 63

6.1 Summary of pre- and post-intervention periods considered for different outcome variables. One year lag was considered between the policy enactment date and the start of the "treated" period, hence we have Sept 2007 instead of Sept 2006 and Apr 2011 instead of Apr 2010. 88
6.2 Model A: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV introduction on IPD incidence for different population subgroups. . 96
6.3 Model B: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV7 introduction on IPD-PCV7 serotypes across population subgroups 96
6.4 Model C: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV7 introduction on non-PCV7 serotypes across population subgroups 101
6.5 Model D: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV13 introduction on PCV13 serotypes across population subgroups . 101
6.6 Model E: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV13 introduction on NV serotypes across population subgroups . . . 103

7.1 Cumulative excess: rate per 100,000, %excess above baseline and %excess above COVID-lab-confirmed, with 95% CrI lower bound (lb) and upper bound (ub). 114
7.2 First date of excess mortality and average daily excess: number of deaths and rate per 100,000 123
7.3 Cumulative excess: number of deaths estimated using synthetic controls (with 95% CrI) and with Poisson regression 125

A.1 Model K: Coefficient estimates for the age-specific model of IPD including Flu, rhinovirus and RSV 181
A.2 Model K: Coefficient standard errors for the age-specific model of IPD including Flu, rhinovirus and RSV 182

List of Abbreviations

AIC Akaike information criterion

AMI average mutual information

AR autoregressive

ARMA autoregressive moving average

ARIMA autoregressive integrated moving average

BSTS bayesian structural time series

CDC US Centers for Disease Control and Prevention

CCM convergent cross mapping

CI confidence interval

CITS controlled interrupted time series

CIM causal impact method

CrI credible interval

DID difference-in-difference

DLM dynamic linear model

EDM empirical dynamical modelling

GP general practitioner

HR hazard ratio

ILI influenza-like-illness

IPD invasive pneumococcal disease

IRR incidence rate ratio

ITS interrupted time series

LRTI lower respiratory tract infections

MSE mean squared error

MSPE mean squared prediction error

NLTS nonlinear time series

NVT non-vaccine type

ONS Office for National Statistics

PHE Public Health England

PCV pneumococcal conjugate vaccines

PPV pneumococcal polysaccharide vaccines

RCGP Royal College of General Practitioners

RSC Research and Surveillance Centre

RSV respiratory syncytial virus

SGSS Second Generation Surveillance System

Chapter 1

Introduction

1.1 History of epidemics

"Epidemic", "pandemic", "contagion", "quarantine" are terms that we have come to know well in recent months due to the COVID-19 emergency. These concepts, however, are by no means unknown to mankind: outbreaks of infectious disease have importantly conditioned the history of human civilizations, as evidenced by numerous historical writings describing plagues and decimated populations over the millennia [81].

The history of epidemics is not recent: epidemics have existed since the first large urban gatherings existed. This is demonstrated, for example, by an Egyptian stele from the 18^{th} dynasty (1403-1365 BC) where a scribe is represented with a leg offended by polio, or the mummy of the pharaoh Ramses (1157 BC), on whose face the pustules of smallpox that probably killed him are still preserved [94]. Smallpox, a virus capable of mowing down over 30% of the population, was successfully eradicated only in the late 1970s thanks to a worldwide vaccination campaign. Before then, it hit the globe for millennia: in Rome, a city that in ancient times comprised a million inhabitants, both the Antonine plague (165-80 AD), which caused 30000 deaths, and the Cyprian plague (in 250-270 AD), which at its peak resulted in 5000 deaths a day, were most likely smallpox epidemics [56].

After that, the bacterium *Yersinia pestis*, causing bubonic plague, has been responsible for the most devastating epidemics recorded in human history: the first one, known as the "Justinian plague", raged from Constantinople to Rome between 541 and 750, killing 50% of the population, with an estimate that historians have calculated of about 30-50 million victims. A second wave of the same scourge 8 centuries later, the "black plague" of which Boccaccio narrates, originated in Asia and eliminated 30-50% of the population (30 million

deaths) [176]. Finally, the third deadliest epidemic occurred just over a century ago: the 1918 Spanish Influenza is considered to have caused 70 to 100 million deaths worldwide [128].

A series of key elements contribute to the outbreak of an epidemic. Zoonosis, or the jump of species from animals to humans, can occur in situations of close proximity to animals and poor sanitation. The domestication of plants and animals developed when, after the last ice age (12-10,000 years ago), man went from small groups of hunter-gatherers to permanent farmers. This change in habits favored the passage of viruses and bacteria to humans, whose immune system was faced with these pathogenic parasites for the first time; animals, on the other hand, had learned to live with them for millennia thanks to a slow mechanism of co-evolution and selection, that had mitigated its virulence and pathogenicity [38].

Low population density is also one of the reasons that explains why the hunter-gatherer bands with groups of 30-50 individuals scattered throughout the Pleistocene period were free of epidemics: with higher contacts between groups of individuals, epidemics became frequent in the era of urban settlements. Similarly, trade and travel contributed to the spread: the silk road was the vehicle of the medieval "black plague" [3], as demonstrated by the origin of the "quarantine" in the ports controlled by Venice in the late 1300s. Ports and ships have been a point of multiplication of the infections on multiple occasions: it was on a ship that in the early 1500s the Spanish conquistadors brought to South America not only weapons, but also infectious agents such as smallpox, measles, salmonella enterica and viral hemorrhagic fever. This explains how the 500 men of Cortés were able to annihilate the large Aztec army of Montezuma. It didn't stop there: in the next 100 years, 22 million deaths followed, reducing the population of Mexico from 25 million people to only 700,000 in 1623 [135].

In this sense, wars can trigger the start of epidemics, with at least two notorious examples: the "Spanish" flu epidemic killed over 70 million individuals in the two years following the First World War, while the Plague of Athens in 430 BC [121] broke out when the city was under siege during the Peloponnesian war. Poor sanitation, food shortages and stress typical of conflicts are likely to have aggravated the spread. At the same time, epidemics can shape the progress of the war and the course of history: Thucydides narrates how Sparta won in such a context of weakness for the Athenian army, imposing oligarchy over democracy for the centuries to come [173]. He also left us a gist of medical knowledge: he noted that those who survive infection are protected from future encounters with the same disease (i.e.

develop immunity), becoming useful members of society for the care of the sick and the various subsistence activities [18].

1.2 Epidemiology and public health of outbreaks

Thucydides has been dubbed the father of scientific history as he gathered evidence and used it to make an analysis of cause and effect, attributing processes to physical reality instead of some god's wrath. However, the history of modern epidemiology and public health typically begins in 1854 in London, where John Snow investigated a cholera outbreak that was happening at the time: his revolutionary idea consisted of using maps and numbers to describe the epidemic [220].

John Snow did not know what caused a disease, as the microscope made its first appearance around that time, nonetheless he speculated that cholera was water borne. He figured out that London neighborhoods were served by different water pumping stations and identified which pumping station was providing water to each neighborhood. He then counted the number of deaths from cholera experienced by those neighborhoods, and divided the number of deaths by the number of houses being served by each pump, obtaining an incidence ratio. His investigation led him to conclude that one particular pumping station, the Broad Street pump, was likely responsible for the majority of cholera deaths in the city. When presenting his results he advised the removal of the pump handle and, despite initial skepticism, his request was granted, ending the cholera outbreak.

Even if today we know that communicable diseases are caused by microbial infectious agents (bacteria, viruses), modern society still faces the important challenge of preventing outbreaks. Like John Snow, modern epidemiologists work as medical detectors to learn the distribution and determinants of disease in a population using non-medical tools. To investigate if a given exposure increases or decreases the likelihood for the outcome of interest, they collect data and formulate a mathematical relationship between exposure and outcome. At this point public health professionals step in: when a statistical relationship between an exposure and an outcome is identified, a control measure on the exposure, a public health intervention, can be implemented to control the outcome. This strategy for disease prevention, on a pragmatic basis, is a viable solution even when the mechanism of how that exposure causes the outcome is unknown, or when it is unknown whether such a mechanism is indeed causal [238].

Compared to John Snow's times, we now have vaccines available for several pathogens, i.e. biological preparations which provide immunity for that specific pathogen. As their administration to individuals who do not have natural immunity is the most effective way to prevent infection, and most infections are transmitted through person-to-person contact or droplets, mass vaccination is the most effective public health intervention to prevent outbreaks in a population. Interestingly, not everyone needs to be vaccinated. When most of the population is immune, the likelihood of disease transmission between non-immune individuals becomes almost null, i.e. the vaccine has induced herd immunity. However, when vaccines are not available, interventions to remove the source of transmission (e.g. John Snow's water pump), or social distancing and quarantine of infected cases to break transmission networks are the only way to prevent and limit outbreaks.

Another major difference compared to John Snow's times concerns data collection. The impact of disease is now quantified not only in terms of mortality, but also in terms of morbidity and effect on the quality of life. Conveniently, most modern countries have in place automated surveillance systems requiring health professionals to report cases of disease included in a list of so-called notifiable diseases to health authorities. Disease surveillance should allow timely identification of changes in the number of cases over time, facilitating early detection of epidemics in the monitored population. This procedure, sometimes referred to as trend analysis, was used for the first time during the 1957-1958 influenza pandemic: change in the national number of deaths for influenza and pneumonia recorded through surveillance was used to estimate the impact of the H3N2 influenza virus [209]; some 60 years later, public health surveillance systems are still considered *"the best weapon to avert epidemics"* [19]. This kind of information is likely to suffer from detection bias, as the more we look for cases (e.g. swab asymptomatic people), the more we are likely to find them, yet its contribution is essential in identifying questions that need further investigation.

In conclusion, in a run towards evidence-based medicine, epidemiology promises to uncover some external truth through data collection and mathematical tools. Detecting the start of an outbreak, quantifying its burden, identifying the contribution of one pathogen and evaluating the effectiveness of an intervention are all research questions of great relevance. The answers to these questions can assist epidemiological understanding of current and past outbreaks to prevent future ones: findings can be employed to inform future public health policies and to help decision-makers in arranging drug stockpiling, vaccination regimes or non-medical interventions.

1.3 Statistical models for epidemic data

As explained in section 1.2, surveillance systems are a major source of data on notifiable diseases. Information routinely collected typically includes longitudinal counts of the number of detected infections, where those detected are most likely those with symptomatic disease. These times series are generally available only at an aggregate level, for example as weekly counts by coarse geographical location, and they might suffer from under-reporting and reporting delay. In fact, data on milder infections, more difficult to diagnose, are more likely to be subject to under-ascertainment and reporting biases.

A first contribution of statistical approaches is then to make sense of this surveillance information by correcting observational biases, to quantify the burden and identify trends over time. A second purpose of statistical approaches is to understand relationships between outcomes and explanatory variables of interest, i.e. to understand the association between disease burden and covariates; or between multiple disease outcomes, i.e. to understand the interaction between different pathogens at the host burden level. This category of analyses may include understanding the effect of public health interventions on the burden of disease.

A final role of statistical modelling is that of using the observed data to reconstruct the unobserved dynamics of the disease of interest. For example, the transmission process underlying an epidemic resulting from the contact between susceptible and infected individuals over time is not observable and so not directly measurable. However, it can be reconstructed using surveillance and other available data (e.g. from ad-hoc surveys), and statistical approaches can assist to combine different data sources [44, 8].

Each of these contexts requires specific statistical models. There are two main types of models: regression-type and mechanistic. The regression-type models, often referred to as phenomenological models, are statistical models expressing the relation between the outcome of interest (e.g. disease incidence over time) and appropriate explanatory variables. They are typically used for the second purpose, and may require simultaneous accounting for the reporting and other observational biases in the data, as in the first aim.

Mechanistic epidemic models, of the SIR (Susceptible, Infected, Recovered) type, approximate the spread of a pathogen in a population [224], describing it as progression through successive disease states. They belong to the family of state-space models [49], where the dependence across disease states is defined as a function of meaningful epidemic parameters. Making inference on these parameters from observed data allows reconstruction of the unob-

served transmission and recovery process, providing estimates of unobserved quantities such as the reproduction number (R_0) and the number of asymptomatic or undetected infections.

1.4 Respiratory infections and co-infections

Among many infectious diseases, this thesis will focus on respiratory infections, in particular lower respiratory tract infections (LRTIs). This term refers to any infection affecting the airways below the larynx, including pneumonia, bronchitis, and bronchiolitis.

LRTIs are responsible for a significant morbidity and mortality worldwide. According to the most recent estimates of Global Burden of Diseases, Injuries, and Risk Factors (GBD) [30], in 2016 LRTIs caused 2.38 million (95% confidence interval (CI) 2.15–2.51) deaths, being the top cause of death in low-income countries, where access to immunisations and antibiotics was limited. LRTIs are a threatening presence in high-income countries too: immunisation is not a solution for newly emerging pathogens, as the SARS-COVID-2 example currently shows us, and the increasing number of drug-resistant pathogens is narrowing the choice of antibiotics that were previously efficacious.

LRTIs can be the result of a broad number of respiratory pathogens, including viruses, bacteria and fungal agents, that can cause both seasonal endemic infections and periodic unpredictable pandemics. The most studied respiratory pathogens, including viruses such as Influenza, Respiratory Syncytial Virus (RSV) or Rhinovirus and bacteria such as *Streptococcus Pneumoniae* or *Staphylococcus Aureus*, are endemic in the human population. Their activity cyclically increases in the autumn and winter months in temperate countries of the world, but thanks to existing immunity their transmission is naturally limited to controlled epidemics. A recent collaboration [30] estimated the contribution of specific pathogens to LRTI burden, concluding that *Streptococcus pneumoniae* was the leading cause of LRTI morbidity and mortality globally, contributing to more deaths than all other aetiologies combined. Infants, people of 65 years of age and above, or individuals with weakened immune systems and other health conditions are typically most at risk for complications. For these groups, immunisation via influenza vaccine and pneumococcal vaccine is recommended in several countries.

Novel respiratory pathogens that have the potential to spread among humans can also emerge: this mostly happens via zoonosis, i.e. when an animal pathogen "jumps" to the human species. Due to non-existing immunity in the human population and absence of vaccines,

transmission can happen very effectively, making containment challenging. Recent examples include the 2009 H1N1 swine flu and three coronaviruses-related illnesses: the severe acute respiratory syndrome (SARS) in 2002, the Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 and the ongoing SARS-CoV-2. These respiratory pathogens have high pandemic potential, causing significant morbidity and mortality, not only among groups considered to be at a higher risk of complications, but also in young, healthy individuals. In addition to public health burden, the restrictions on travel and trade imposed to achieve containment are likely to have significant economic, social, and political consequences.

In the interest of reducing burden and pandemic preparedness, therefore, understanding the burden of LRTIs, how different LRTI-causing pathogens might interact, and the effect of public health interventions such as vaccination on the burden remain crucial subjects of much research.

1.5 Challenges in estimating LRTI burden

Two similarities in LRTI-causing pathogens result in challenges in identifying the pathogen responsible for an LRTI and estimating its burden on healthcare: similar, largely unobservable, transmission routes, via droplets and aerosols [196]; and non-specificity of symptoms of LRTI [22]. In the first place, since most symptoms of respiratory infection are mild, many individuals will not access any healthcare and hence will not appear in surveillance data. For the minority of cases where infection manifests as a severe LRTI, whose symptoms include fever, a severe cough, difficulty breathing or chest pain, the diagnostic process is generally limited to a physical exam through a stethoscope by your General Practitioner (GP) and chest X-rays. It is rare that blood or mucus samples are taken to test for specific bacteria and viruses. As symptoms are not pathogen-specific, the pathogen responsible for infection is rarely detected in routine clinical practice, resulting in substantial under-reporting of disease incidence. It is therefore a challenge to disentangle the burden on healthcare contacts (GP consultations, hospital admissions) and mortality attributable to different LRTI-causing pathogens.

However, integrating routine surveillance data on a non-specific LRTI outcome, e.g. GP consultations for influenza-like-illness (ILI), with some accurate measure of pathogen-specific incidence is often a possibility. A number of ad-hoc studies have been implemented across countries to complement the missing information. For example, a sentinel virological

testing scheme in England, run by the Royal College of General Practitioners [34], routinely tests a proportion of patients consulting for ILI symptoms for a set of respiratory viruses, so that a proxy for seasonal circulation of viruses such as influenza, RSV and rhinovirus can be inferred.

A second challenge concerns understanding how different LRTI-causing pathogens might interact in affecting transmission and host susceptibility when co-infecting the host. “Coinfection" is broadly defined as the condition where a host is *simultaneously* infected by multiple pathogen species or strains, but the term is more often used to imply the presence of *interaction* between infectious agents. They might act synergistically, leading to a superinfection (i.e. the burden of one pathogen is enhanced by the other), or they could compete, with one of them suppressing the second (antagonist interaction) [88]. The host’s immunity also appears to play a key role in coinfection, as an over-regulation of the inflammatory response might exacerbate the damages due to the infectious agents, leading to increased disease severity [73].

The particular role of influenza virus and its underpinning mechanisms in LRTIs have been extensively investigated by biologists since the 1918 pandemic [144, 189], as post-mortem examinations at the time revealed the presence of bacterial infection in the lungs of many influenza-infected individuals [148]. A synergistic interplay between influenza and *Streptococcus pneumoniae*, graphically represented in Figure 1.1, is considered to be one of the main drivers of severity [201]: this has been validated in animal models [132], however the underlying biological mechanism is still poorly understood and empirical measures remain difficult and expensive to obtain in humans [43]. Severity is not the only possible consequence of coinfection: one infection might precede the other, enhancing susceptibility to the second one by impairing immune response and facilitating the entrance of pathogens that are usually cleared by the natural defense system. For example, Shrestha et al. [189] found increased susceptibility to bacterial infections after influenza.

Longitudinal cohort studies carrying out individual-level testing for multiple pathogens are one possibility for improving the current incomplete understanding of the biological interplay between different pathogens. However, such studies are expensive and time-consuming, with only a few available to understand, for example, carriage and transmission. In the case of *S.Pneumoniae*, nasal carriage, an initial asymptomatic stage, has been identified as a necessary but not sufficient condition for the development of disease [191]: most people will not develop systemic disease after carrying pneumococci, and especially in early childhood this condition has been observed to be particularly common and long-lasting [120]. Car-

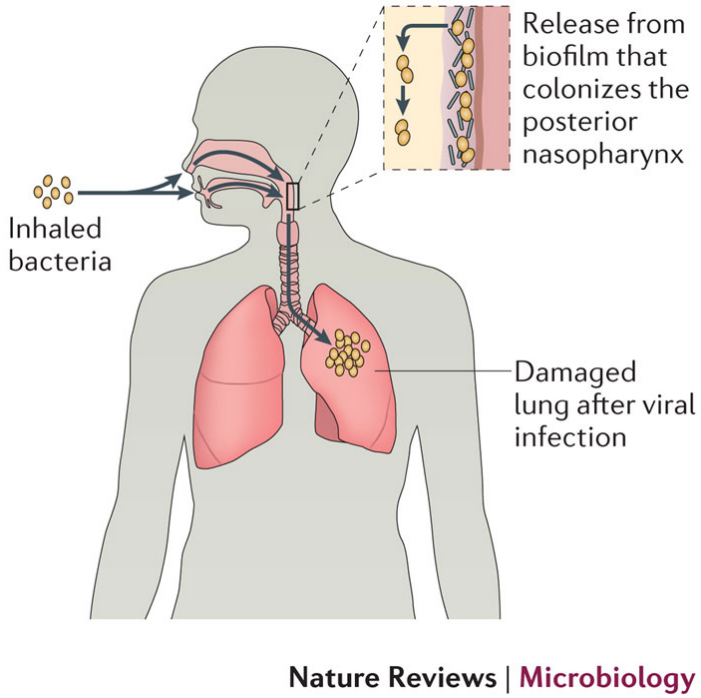


Fig. 1.1 Process of bacterial progression from upper to lower respiratory tract triggered by viral infection [132]

riage studies have been performed on children across several countries and helped unravel dynamics of these infection stages [136]. However, modelling pneumococcal dynamics in carriage studies, and in particular understanding the interaction with other pathogens such as influenza, raises several challenges. These include intermittent observation schemes, potential informativeness of observation times, and lack of biological understanding limiting the formulation of transmission models without strong assumptions [162].

An alternative is to use phenomenological models for population-level aggregate data on multiple pathogens to attempt to disentangle their interaction from common confounders and drivers such as meteorological processes. However, such an approach brings its own challenges, the biggest one being the observational nature of the available information, as anticipated in section 1.3. The absence of randomisation, and the presence of confounding, do not allow making the causal statements that might be needed for the understanding of disease determinants. However, in the field of burden estimation and public health interventions, improvements to currently and routinely employed approaches to attribution in the presence of confounding (often limited to conclusions of association only) can be envisaged through the estimation of what is known as a "counterfactual". As David Lewis puts it, "We think of a cause as something that makes a difference, and the difference it makes must be a difference

from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well" [116]. This leads to another challenge in both estimating LRTI burden and evaluating the effect of public health interventions on LRTI burden: how to estimate such a counterfactual, since it can never be observed?

In this thesis, motivated by substantive problems in understanding LRTI burden, we critically appraise currently used methodologies, identifying shortcomings and proposing improved approaches. We give particular emphasis to phenomenological time series models for burden estimation and intervention evaluation, where the common theme is the estimation of counterfactuals in the presence of observed and unobserved confounding: a challenging task.

1.6 Aims of the thesis

This thesis proposes and investigates statistical models to quantify burden of respiratory disease and evaluate effectiveness of public health interventions while accounting for the challenges posed by surveillance data (section 1.5), and applies them to different study designs. A variety of data on respiratory disease incidence, typically of longitudinal nature, are used to achieve these aims.

First, in chapter 2, a reconstruction of pneumococcal disease progression from asymptomatic to severe phase is attempted. Data from a cohort study, which enrolled children in a developing country and actively ascertained disease status and measured individual risk factors, provided individual-level information about pneumococcal carriage and LRTI. We model the stages of disease from asymptomatic to invasive through a multi-state model, accounting for viral circulation and climatic factors. Exploring the factors triggering pneumococcal disease progression in a vaccine-free setting can improve our understanding of pneumococcal disease dynamics and the interaction with viruses such as influenza and RSV.

Next, in chapter 3 we review the literature on phenomenological models, focussing on aggregate outcome data. With the scope of clarifying the role of viral coinfection and meteorological conditions in the development of invasive pneumococcal disease (IPD), in chapter 4 we propose a novel multivariate linear regression to quantify the contribution of respiratory viruses on the incidence of severe pneumococcal infections by age group. The approach integrates GP consultations for influenza-like-illness with RCGP viral positivity to

obtain proxies of viral circulation; whereas IPD incidence data are obtained from the national surveillance system. The results provide crucial evidence to support decisions on antibiotic stockpiles in view of a future influenza pandemic.

Chapter 5 presents statistical methods for the evaluation of interventions, with a focus on public health examples, and chapter 6 inspects the changes in pneumococcal disease incidence following the introduction of pneumococcal conjugate vaccines. Using serotype-specific information on IPD counts, we quantify the impact of pneumococcal vaccine introduction using interrupted time series and Bayesian structural time series methods. We disentangle the contribution of serotype replacement across age groups by separately estimating the reduction in vaccine-targeted serotypes and the increase in non-vaccine types. Measures of effectiveness of the recently implemented vaccination strategies and quantification of their side effects are fundamental to inform health policy, in particular to support future decisions on the introduction of higher-valency vaccines.

In chapter 7, we quantify the excess all-cause mortality by age and by region during the COVID-19 pandemic period in England. A counterfactual is estimated including control time series through a dynamic regression model, using again a Bayesian structural time series approach.

Finally, chapter 8 presents an alternative empirical dynamical modelling (EDM) methodology to quantify causal interactions between time series. The method is applied to the IPD and influenza time series data used in chapter 4 to understand their interaction.

Chapter 9 wraps up with some final discussion.

Chapter 2

Individual-level dynamics of disease

2.1 Introduction

Longitudinal studies are occasionally run on population subgroups to track disease evolution, especially for pathogens characterised by an asymptomatic carriage phase, such as the pneumococcus. When available, individual-level data can be analysed to provide useful insights about the viral-bacterial coinfection process.

In this chapter we use data from a cohort of children followed up for two years, to infer pneumococcal disease incidence as a function of seasonality and viral circulation. The information available allows modelling within-subject pneumococcal dynamics over time, as the asymptomatic carriage state (colonisation) with the pneumococcus is ascertained at predefined times and LRTIs are identified due to self-reporting to healthcare due to the severity of symptoms.

Research questions in this chapter are: do viral circulation, wet season and high temperature facilitate or protect from acquisition of pneumococcal carriage? Is clearance affected to the same extent? A question regarding the progression from carriage to LRTI is also of interest: do these factors trigger the occurrence of LRTIs? In other words, how many of these are bacterial infections secondary to influenza? These questions can be addressed by approximating the natural history of pneumococcal disease through a multistate model, as we elaborate in what follows [138].

2.2 Multistate models

Multistate models allow the disease history of an individual to be expressed as a series of states. Each state represents the possible disease condition at a given time, and the movement through states indicates disease progression [4]. The simplest example of a multistate model is the survival model, which only has two states, alive and dead. After defining the set of allowed transitions from each state, we are interested in estimating the time spent in each state before the transition happens, the probability of moving to a specific state, and the proportion of individuals who are in each state at any given time.

Ideally, the disease history would be entirely described by observed transition times of each individual between states, however in many practical applications information is not available in such detail, and only the current state can be observed. This is the case for chronic or asymptomatic conditions, when the clinical state is typically ascertained at predefined follow-up visits [95]. Reconstruction of the movement between states can be considerably more challenging when based on such panel data, as evolution of the process is effectively unknown between the observation times.

In the next sections we introduce the notation and the key features of the most commonly used multistate models, delineate the structure for likelihood-based inference in the general case, and finally focus on inference from panel data.

2.2.1 Notation and assumptions

Consider a system characterised by a finite number of states $s \in S = \{1,2,3,\ldots,k\}$. S is called the state space. Define $\{Y_i(t), t \geq 0\}$ to be a random variable denoting the state occupied by subject i at time t. Also, denote by $H(\tau)$ the clinical history up to time τ, i.e. $H(\tau) = \{Y_t, 0 \leq t < \tau\}$. The process is uniquely characterised by the transition intensity function $q_{r,s}(\tau, H(\tau))$, that defines the set of possible states the process may move to when it leaves state r

$$q_{r,s}(\tau, H(\tau)) = \lim_{dt \to 0} \frac{Pr\{Y(\tau+dt) = s \mid Y(\tau) = r, H(\tau)\}}{dt} \tag{2.1}$$

where (r,s) identifies any pair of states, and $q_{r,s}(\tau, H(\tau))$ is the element in position (r,s) of the $k \times k$ transition intensity matrix $Q(\tau, H(\tau))$ [36].

In formula 2.1, the probability distribution of the future state $Y(\tau + dt)$ is a function of both the present state $Y(\tau)$ and the past disease history $H(\tau)$. The process is said to be Markovian, or memoryless, if such a probability does not depend on disease history. Transition intensities are functions of the current state r and time τ only: $q_{r,s}(\tau, H(\tau)) = q_{r,s}(\tau)$.

A further simplification in a Markov model is to assume that the transition intensities are constant over time, i.e the process is time-homogeneous: $q_{r,s}(\tau) = q_{r,s}$. To relax the assumption of time-homogeneity, some time-varying covariates z_t can be incorporated in the model to make it more flexible. The transition intensities are usually modelled as a function of such time-dependent variables under the assumption of proportional hazards:

$$q_{r,s}(z_t) = q_{r,s}(0) \exp(\beta_{rs}^T z_t) \tag{2.2}$$

where $q_{r,s}(0)$ represents the nonparametric baseline hazard and β_{rs}^T are the transition-specific regression coefficients.

From the specification of the transition intensities $q_{r,s}$, the one-step transition probabilities of moving to state s when starting from r at two arbitrary points in time are defined as

$$p_{r,s}(\tau, \tau + dt) = P(Y(\tau + dt) = s \mid Y(\tau) = r), \forall \tau \in (0, \infty). \tag{2.3}$$

Transition probabilities can be calculated by solving the Kolmogorov differential equation: $P'(dt) = P(t)Q$. where Q is the matrix whose generic element is $q_{r,s}$ and P is the matrix with entries $p_{r,s}$. This equation has a solution through matrix exponentiation: $P(dt) = e^{Qdt}$, however it becomes computationally intractable as the number of states and transitions increases [76].

2.2.2 Estimating transitions using panel data

In longitudinal cohort studies of the type considered here, the exact time of transition will not be observed. This panel data structure is represented in Figure 2.1.

Denoting by $t_{ij},\ j = 1, 2, 3, \ldots, J$ the sequence of time points where the state Y_{ij} for individual i is observed, we are interested in reconstructing the entire disease history, i.e. estimating the transition intensity matrix Q. In order to make inference from such incomplete information, some assumptions are usually imposed to obtain a tractable likelihood. Several

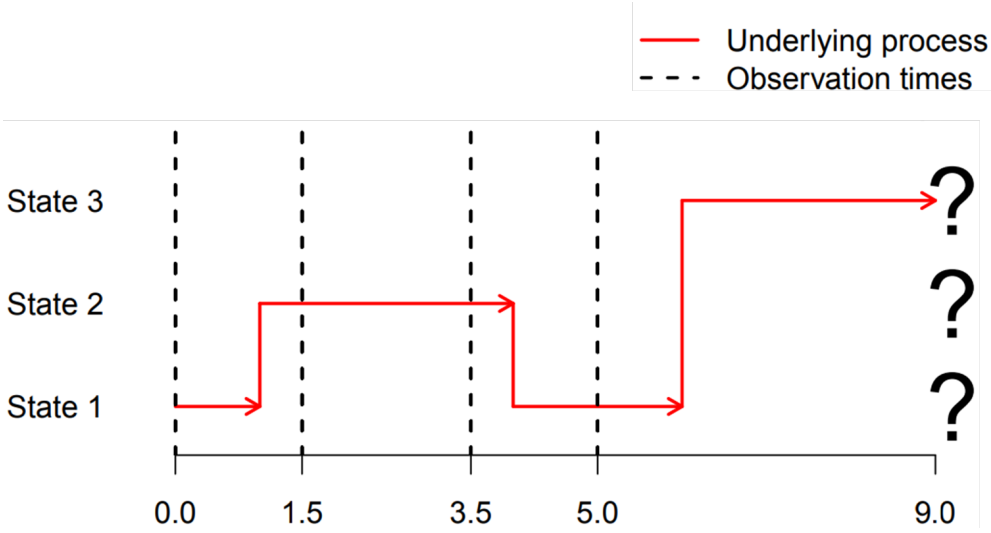


Fig. 2.1 Example of process observed under a panel data scheme

.

estimation methods are based on the assumption that times t_{ij} at which observations happen are not informative about the state observed [32]. Among them, Kalbfleisch and Lawless [107] proposed a method to estimate transition intensities from the transition probabilities of the discrete-time Markov chain embedded in the Markov process studied. Namely, the likelihood function for the entire process, based on the series of Y_{ij}, i.e. the state in which individual i is at the observation time j, can be written as:

$$L(Q) = \prod_{i=1}^{n} \prod_{j=1}^{J} L_{i,j} = \prod_{i=1}^{n} \prod_{j=1}^{J} p_{(Y_{ij},Y_{i,j+1})}(t_{i,j+1} - t_{ij}), \tag{2.4}$$

i.e. the product over all the individuals and observation times of the probabilities of transition between the observed states.

2.2.3 Mixture of observed and unobserved transition times

If a state v is exactly observed, thanks to Markovianity and time-homogeneity, the Chapman-Kolmogorov equation allows us to write:

$$p_{r,v}(t_{i,j+1} - t_{i,j}) = \sum_{s \neq v} p_{(r,s)}(t_{i,j+1} - t_{i,j-})q_{s,v} \tag{2.5}$$

We are factorising the transition probability $p_{r,v}$ in two components, based on the fact that the state immediately before transitioning to v is unknown. We consider the probability of

moving from r to any state s, and the instantaneous transition from s to v, summing over all the possible states $s \neq v$.

The same patient might be characterised by a mixed observation scheme, with some states only observed at predefined visits, and other states whose transition is exactly observed. To make efficient use of all the information, these heterogeneous transitions are combined in a single model, so that the likelihood $L(Q)$ will be a mixture of terms described in equations (2.4) and (2.5), i.e.

$$L(Q) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_{r,s\neq v}(t_{i,j+1} - t_{i,j}) * \Big[\sum_{s\neq v} p_{(r,s)}(t_{i,j+1} - t_{i,j-})q_{s,v} \Big]. \tag{2.6}$$

2.3 Application to pneumococcal disease

An application of multistate models to pneumococcal disease dynamics is considered in this section. We first present a brief review of the literature and the data at hand. We then discuss modeling considerations, that is, how to build a sensible model for the problem of interest. At last, we present and discuss results on the selected models.

2.3.1 Previous multistate models for S. pneumoniae

Several longitudinal studies have been established to inspect pneumococcal dynamics, in particular the unobservable processes of colonisation acquisition and clearance: samples have been typically collected at regular pre-defined time intervals (e.g. every month), and transition rates have been estimated from panel data under continuous-time Markov models [138, 74, 91, 2]. After the introduction of the Pneumococcal Conjugate Vaccines (PCV) in several countries, much attention has been paid to the behaviour of different serotypes, with disease status more often defined by distinguishing colonisation with vaccine (VT) or not vaccine type (NVT)[218]: importantly, Auranen et al. [5] and Mehtälä et al. [137] have investigated competition among serotypes. However, to our knowledge, no study has combined information about upper respiratory tract carriage and LRTI. An improved understanding of probability and timing of disease progression with respect to seasonality would shed light on

the coinfection process.

2.3.2 Source of data

The pneumococcal data of interest refer to the study conducted in the refugee camp of Maela, on the Thai-Myanmar border. We thank Prof Paul Turner and Prof Claudia Turner for sharing the data. A total of 999 children have been recruited at birth, from October 2007 to November 2008, and followed for 24 months. A nasopharyngeal swab has been taken and tested monthly by healthcare professionals of US Centers for Disease Control and Prevention (CDC) and Shoklo Malaria Research Unit. Serotyping has also been performed on a portion of the samples, but that information will not be included in our analysis.

This study is nested within a longitudinal study designed to establish the epidemiology and etiology of pneumonia in children [211]. As primary health services were not provided to refugees outside the camp, all the clinical cases were referred to a field hospital and two outpatient clinics. This situation enables us to reconstruct a complete disease history for each child, including both the asymptomatic and symptomatic phases.

Details about meteorological conditions in the camp, such as monthly rainfall, average, minimum and maximum temperature, have also been recorded for the entire study period (October 2007 to October 2010). The same researchers have also measured viral positivity in the pneumonia cases associated with a suspected upper respiratory tract viral illness (symptoms like cough, sneeze, runny nose, strep throat) [212]. Figure 2.3 summarises the dynamics of the described phenomena at the Maela camp over time.

Finally, we gather information about viral positivity collected through the ILI sentinel surveillance (in 11 sites throughout the country) established in 2004 by the National Institute of Health in Thailand (NIH), in collaboration with the CDC. While the system was intended to monitor the frequency of influenza virus infection and describe seasonality, testing for 6 additional respiratory viruses is also performed [23]. Information is presented in Figure 2.4.

2.3.3 Analysis strategy

When using a multistate model to describe disease dynamics over time, the state structure must be defined in the first place. Assuming that every child is born healthy, our model is made of three states which define pneumococcal disease progression: uncolonised, colonised

and infected with pneumonia.

In the next stage, allowed transitions must be identified: we are interested in modelling the time to pneumococcal carriage acquisition from uncolonised. Once colonisation has been acquired, the stay in this state is temporary: the body immune response might clear the bacteria, or it might lead to a respiratory infection, generally pneumonia. Therefore, we assume that, instantaneously, individuals can move from uncolonised to colonised and move back, while the progression to pneumonia necessarily happens from carriage. Finally, we consider pneumonia to be an absorbing state, as we are interested in the etiology of pneumonia but not strictly in recurrent pneumonias, that could have different etiological mechanisms (Figure 2.2).

Lastly, proposing realistic model assumptions is necessary. In the highlighted situation it is reasonable to assume that our model is Markovian, i.e. the next disease stage only depends on the current clinical condition, regardless of the past history. From available medical knowledge, we are confident that previous exposure to carriage does not provide immunity to subsequent carriage nor facilitate it, and presumably chances of developing pneumonia do not depend on duration of carriage. Markovianity facilitates the inference and also allows for an intuitive graphical understanding of the model.

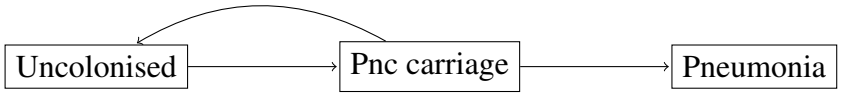


Fig. 2.2 Transitions between states that are instantaneously possible

As described in section 2.2.3, our data are a mixture of observed and unobserved transition times: colonisation can be detected only when a nasopharyngeal sample is taken and cultured, hence information on the presence of *S.Pneumoniae* was actively looked for through a program of monthly swabbing. Information is available on the prevalent condition, but acquisition and clearance times are unobservable. However, as times of visits were predefined, and remembering that carriage is an asymptomatic condition, we are confident these observation points are uninformative about the actual presence or absence of bacteria. On the other hand, pneumonia is a severe condition, especially in infants. We assume the transition from carriage to pneumonia is exactly observed, as parents would immediately seek healthcare when the symptoms appear. Therefore, the likelihood takes the mixed form expressed in equation (2.6). Maximum likelihood estimates for the transition intensities are

obtained using the *msm* R package.

Finally, covariates about national surveillance for viral infections and about weather conditions in the Maela camp are sequentially added to the model, using the Cox regression structure presented in formula (2.2). The rationale for such choice is that we want to investigate if pneumococcal disease, and in particular the progression from colonisation to pneumonia, is in any way influenced by Influenza or RSV circulation, after taking into account shared drivers such as temperature variations. Formal model comparison is performed using likelihood ratio tests, starting from the model with no covariates and adding the most significant predictor at each stage. Minimum temperature is included as a continuous variable, whereas indicator variables, as a proxy for viral circulation, are defined as “Flu season" if at least 20% of the tested samples resulted positive to the virus, and “RSV season" if at least 5% of the samples tested positive to it (considering that RSV positivity reaches a maximum of 15% in the general population). Similarly, a dichotomous indicator for wet season is defined according to the Köppen climate classification **??** for tropical climates: any month when average precipitation is below 60 millimetres is considered dry.

2.4 Results

2.4.1 Descriptive analysis

A total of 740 subjects are included in the analysis, i.e. all the children who had at least one swab. 15282 records are available, with a median number of 25 clinical observations per child (IQR 13-26, min 2, max 37). 14197 records refer to regular swabs, of which 9895 resulted positive (69.7%). 1085 observations identify pneumonia episodes, of which 488 refer to a first pneumonia episode, while the remaining 597 concern recurrent pneumonias in the same subjects (up to 12 per child over 24 months).

Figure 2.3 summarises weather and disease conditions for the Maela cohort. In the first panel, meteorological conditions at the camp during the study period are displayed: the wet season repeats cyclically between April and September, with rainfall amounts not too dissimilar across years. The dry season, between October and March, is instead characterised by lowest minimum temperatures, with monthly averages below 20. The second panel presents carriage prevalence: except for the first semester of the study, when children were still being recruited, carriage prevalence fluctuates around 60% over the entire observed

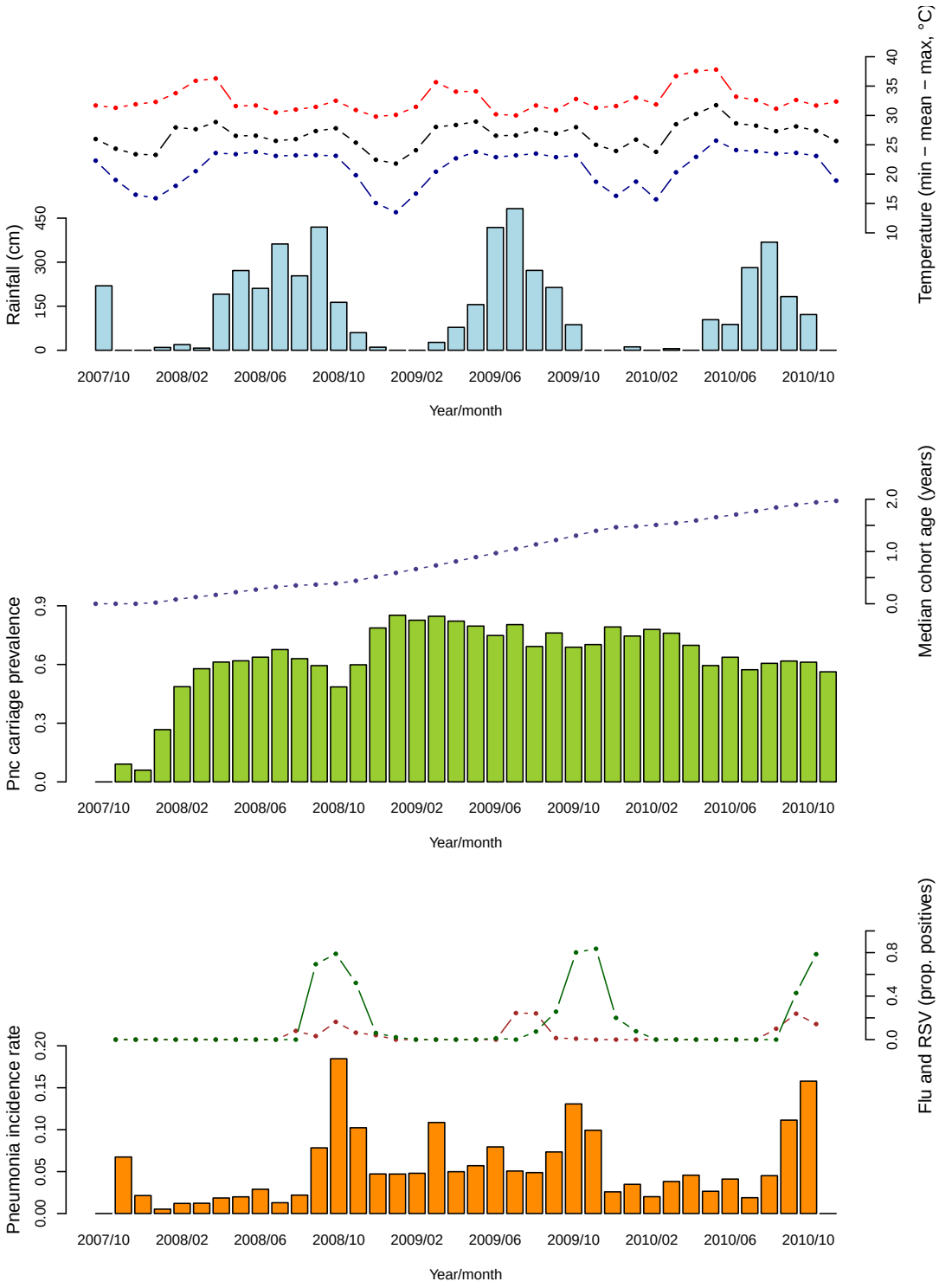


Fig. 2.3 Weather conditions, carriage prevalence, median cohort age, pneumonia incidence and viral circulation.

period, with no obvious seasonal patterns except for some reduction between the end of the wet season and the start of the dry one (August-November). Finally, the third panel presents incidence of pneumonia, that peaks yearly between September and November. High prevalences of RSV and Influenza are identified at the same time, however we consider this information to be biased, as testing only happens in correspondence of these acute infections.

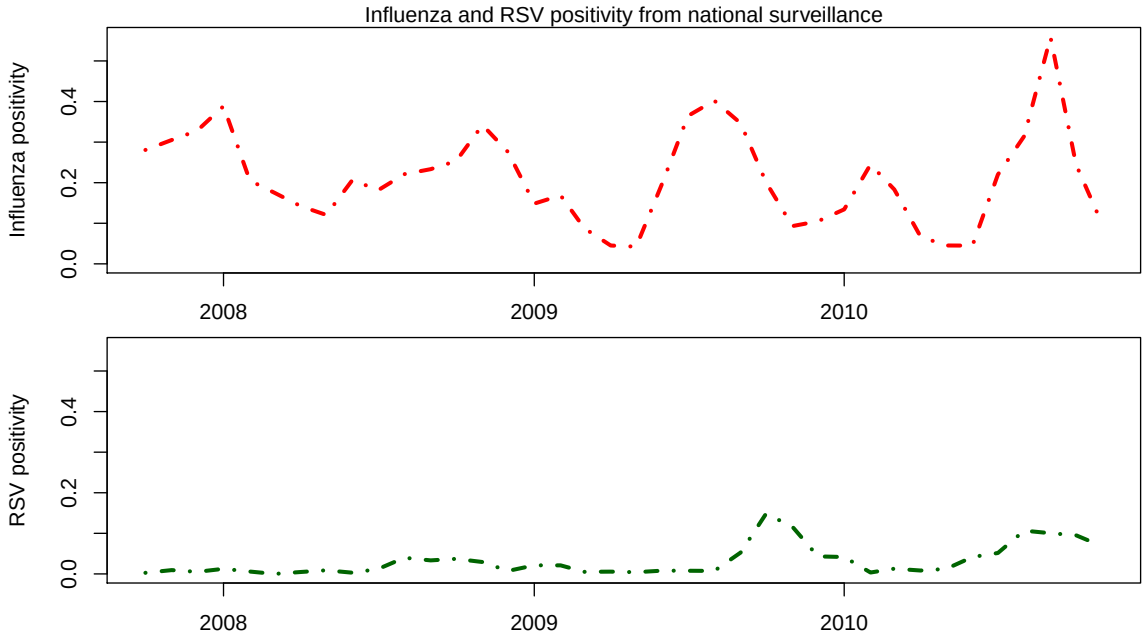


Fig. 2.4 Viral positivity obtained from the national surveillance system of Thailand.

Finally, Figure 2.4 presents viral positivity obtained from the national surveillance system: the timing of influenza seems to vary across seasons, with the peak occurring between October and December in 2007 and 2008, whereas an earlier peak, around August, can be observed for 2009 and 2010, with an intermediate wave in spring, realistically the second 2009 flu pandemic wave. RSV positivity is not as marked in the first two seasons (possibly due to recent introduction of its viral testing), however some epidemic behaviour can be identified around October-November in 2009 and around August-September in 2010.

2.4.2 Multistate model

After limiting the follow-up to the first pneumonia episode, we are considering a cumulative time of 740 person-years, ending with a pneumonia episode for 488 children (66% of the cohort), while the remaining 252 subjects are censored for the purposes of our analysis. 8628

swabs were tested during the study time, with positivity to pneumococcal carriage in 67.2% of instances.

Observed transitions are presented in table 7.3: having multiple observations of carriage in a sequence looks very common, but also staying uncolonised at subsequent occasions is common, while clearance is less frequent. Transition to pneumonia is more common from carriage than from uncolonised, and an intermediate unobserved transition via carriage state is assumed for the latter. Importantly, this table does not provide information about time of stay in each state.

	Uncolonised	Pnc carriage	Pneumonia
Uncolonised	1183	1440	121
Pnc carriage	911	4354	367

Table 2.1 Observed transitions across states from the dataset considered.

	Uncolonised	Pnc carriage	Pneumonia
Uncolonised	-1.004 (-1.069,-0.942)	1.004 (0.942, 1.069)	0
Pnc carriage	0.304 (0.281, 0.329)	-0.382 (-0.408,-0.357)	0.077 (0.071, 0.084)
Pneumonia	0	0	0

Table 2.2 Estimated transition intensities (and 95% confidence intervals) for the multistate model in Figure 2.2.

The multistate model in Figure 2.2 is initially fitted without covariates in order to describe the general disease dynamic. Time is expressed in months. Maximum likelihood estimates for the transition intensities are shown in table 2.2. This first inference suggests that the instantaneous hazard of carriage acquisition is quite high, i.e. an infant becomes colonised on average within one month from birth. However, after onset of carriage, children are 4 times more likely to clear colonisation than to develop pneumonia: probability of transitioning to the pneumonia state is 0.201 (95% CI 0.184, 0.224), compared to probability of clearing carriage being 0.798 (95% CI 0.776, 0.816). The rate of pneumococcal clearance is low, suggesting that, after pneumococcus is installed in the upper respiratory tract, a long permanence in that condition is expected. Finally, rate of pneumonia onset is even lower, indicating that carriage can persist for over one year before pneumonia is developed.

An estimated mean sojourn time per visit, by state but not specific to the state of destination, is presented in the first column of table 2.3. The other two columns summarise, over

	Sojourn duration (months)		Number of visits		Length of stay (months)	
	Mean	95% CI	Mean	95% CI	Total	95% CI
Uncolonised	1.00	(0.94; 1.06)	2.98	(2.73; 3.25)	3.91	(3.68; 4.15)
Pnc carriage	2.62	(2.45; 2.80)	3.92	(3.67; 4.18)	9.80	(9.27; 10.30)
Pneumonia	-	-	0.76	(0.73; 0.79)	-	-

Table 2.3 Estimated mean sojourn for each state visit, expected number of visits to each state in the two years of follow-up, and forecast total length of time spent in each state in the two years of follow-up for the multistate model in Figure 2.2.

the two-year follow-up, the mean number of visits to each state and the total length of stay in each stay: after being born uncolonised, children on average clear carriage another two times, spending under four months in total in this condition. Conversely, permanence in the carriage state sums up to almost ten months on average before pneumonia is developed, with an average of four visits.

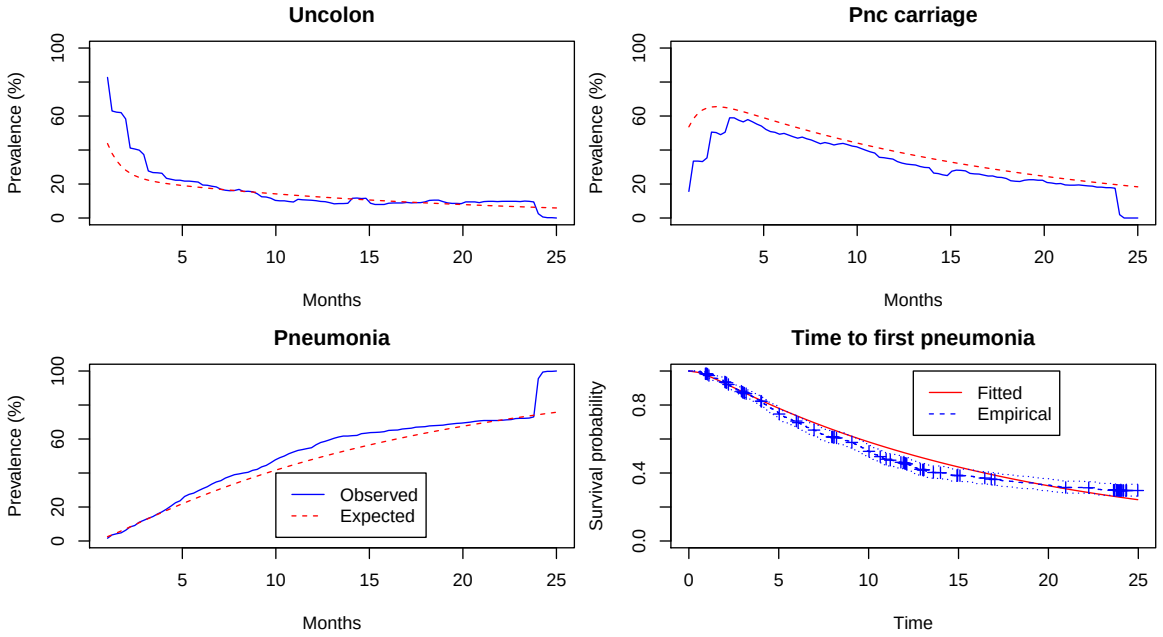


Fig. 2.5 In the first three panels: observed and estimated prevalence across different states over study time. In the last panel: survival function for time to pneumonia.

Figure 2.5 summarises how prevalence of each state varies over the study time, with the first panels comparing the observed numbers of individuals occupying each state and the numbers estimated from the fitted model. The proportion of uncolonised children reasonably diminishes as they progress towards carriage and then pneumonia. For example, only 20%

of children are still uncolonised at four months from birth, while 60% of them have some carriage and the remaining 20% has already developed pneumonia. These plots also provide an approximate indication of the goodness of fit of our model: from a graphical inspection we can see how distances between the curves are quite small in all panels, suggesting a good fit.

Finally, the fourth panel shows the survival function for transition from uncolonised to pneumonia: from the survival curve we can see that, after taking censoring into account, the median time to pneumonia onset is 10.61 months (95% CI 9.89-11.86). We can also see that 25% of children develop pneumonia in the first five months of life, and just over 30% of them do not develop any pneumonia during the study time. Aiming to assess goodness of fit once again, we are reassured to see that our model prediction closely follows the estimated Kaplan-Meier survival curve.

The results presented above are obtained under the assumption that transition intensities are constant over the entire study period. Since this assumption may not hold, we allow calendar-time variations by introducing some time-dependent covariates in the model. We fit the models sequentially, adding one covariate at a time and testing its significance compared to the model fitted at the previous stage using a likelihood ratio test. Table 2.4 summarises the model selection steps:

	Selected var	-2 log LR	df	p
One var	Flu season	66.1793	3	≤ 0.001
Two vars	Wet season	50.5304	3	≤ 0.001
Three vars	Min temperature	61.3549	3	≤ 0.001
Four vars	RSV season	14.2728	3	0.0026

Table 2.4 Model selection steps: LR test for addition of covariates to the model in Figure 2.2.

The selected model includes indicator variables for influenza, RSV and wet seasons, and the continuous measurements of minimum temperature. The estimated hazard ratios (HR) are presented in table 2.5. Firstly, the acquisition of carriage does not seem to be affected by any covariate. Secondly, chances of clearing carriage are significantly lower during wet season (HR=0.707, 95% CI 0.515;0.972) and in months with higher overnight temperature (HR=1.078, 95% CI 1.026;1.133), but they are not affected by viral circulation. Thirdly, all the considered covariates affect the chances of leaving carriage state to develop pneumonia: the predominant factor is the wet season (HR=5.792, 95% CI 3.723; 9.011), followed by influenza circulation (HR=1.67, 95% CI 1.353; 2.062) and RSV season (HR=1.47, 95% CI

	Flu season		Wet season		Min temperature		RSV season	
	HR	CI	HR	CI	HR	CI	HR	CI
Uncolonised - Pnc carriage	0.901	(0.783; 1.037)	0.802	(0.626; 1.027)	0.99	(0.954; 1.028)	0.934	(0.744; 1.173)
Pnc carriage - Uncolonised	1.107	(0.933; 1.313)	0.707	(0.514; 0.972)	1.078	(1.026; 1.133)	1.279	(0.989; 1.655)
Pnc carriage - Pneumonia	1.667	(1.353; 2.062)	5.792	(3.723; 9.011)	0.818	(0.771; 0.869)	1.47	(1.101; 1.964)

Table 2.5 Estimated HRs (95% CI) for the covariates included in the selected model

1.101; 1.964) and lower temperature (HR=0.818, 95% CI 0.771; 0.869).

Finally, Table 2.6 shows a comparison between instantaneous transition rates and length of stay by state across seasons. In particular we compare a period of wet season with high Influenza circulation with a dry time of the year when influenza isn't circulating. As expected, humidity and viral circulation impact the risk of developing pneumonia, hence the main difference between the two scenarios refers to the instantaneous transition from carriage to pneumonia state, almost 10 times bigger. However, confidence intervals are quite wide. Correspondingly, the estimated duration of carriage is cut by a factor of 10 in the wet season, while the time spent in the uncolonised state is slightly longer in the wet season (approx 27 days instead of 20) compared to the dry one.

2.5 Discussion

These results give us an example of how partial information can be used to reconstruct pneumococcal disease dynamics and assess the impact of potential risk factors. Importantly, our results highlight how wet season and low temperatures are important factors both for persistence of pneumococcal colonisation and for its progression to the lower respiratory tract. Even more importantly, our model estimates that the risk of developing pneumonia is 67% higher during the influenza season, and 47% higher during the RSV season, after taking into account meteorological variations. This result still stands out after considering the factors altogether in the same model, and uncertainty around the coefficient estimates is quite small despite estimating fifteen parameters.

Lack of typical seasonality for influenza and RSV is an advantage to our model, as it allows us to better disentangle the contribution of meteorological factors from viral impact. However, RSV contribution might be diluted due to poor detection: ILI case definition and other aspects of influenza virus surveillance may not be optimally suited for other respiratory viruses. Furthermore, we are using positivity in ILI patients in the general Thai population, as we lack age-specific information, but we can imagine RSV incidence to be higher in children.

Our modelling choices might have some limitations: firstly, we are assuming that our variables are observed without error, whereas for example swab tests could suffer from some false positives or false negatives. A hidden Markov model would allow specifying misclassification probabilities. Further, we are considering temperature as a continuous

variable, implying linearity of effects, and cut-offs to define viral seasons were chosen based on data inspection. Other predictors have been deliberately left out: age, for example, could have been an important factor if we assumed immunity was built over time. Similarly, details about household size and structure, information about pregnancy and delivery, location of the house, presence of animals etc have not been investigated, for the sake of model interpretability and computational power. Finally, more work would be needed to assess the adequacy of assumptions in our model: proportionality of hazards does not always hold, and the time-homogeneity assumption may be violated as well. However, alternative models do not have easy implementation for panel data.

More generally, identification of the etiologic agents of pneumonia remains a diagnostic challenge due to the difficulty in obtaining adequate samples for culture from the infection site [222], and this has hampered understanding of the relationship between pneumococcal carriage and pneumonia. We cannot rule out whether other pathogens were responsible for some of the observed episodes, however pneumococcus is recognized as the most important cause of bacterial pneumonia in children aged less than 5 years [159]. Further, generalization of these results to other populations might not be straightforward, as it has been previously observed that prevalence of pneumococcal carriage is much higher in developing countries, and the first acquisition happens at a much younger age compared with the industrialized settings [141].

Finally, information about serotypes could be modelled in the future, as it has been speculated that colonisation with some types is more likely to lead to severe disease, or be characterised by longer colonisation phases [120].

Covariates		Uncolonised	Pnc carriage	Pneumonia	Length of stay (CI)
Wet seas=0,	Uncolonised	-1.5287 (-3.042,-0.768)	1.529 (0.768, 3.042)	0	0.65 (0.33; 1.30)
flu seas=0	Pnc.carriage	0.072 (0.029,0.180)	-1.164 (-3.067,-0.442)	1.092 (0.390, 3.059)	0.86 (0.33; 2.26)
Wet seas=1,	Uncolonised	-1.107 (-2.557,-0.479)	1.107 (0.479, 2.557)	0	0.90 (0.39; 2.09)
flu seas=1	Pnc.carriage	0.056 (0.018, 0.174)	-10.608 (-39.67,-2.836)	10.552 (2.802,39.74)	0.09 (0.03; 0.35)

Table 2.6 Estimated transition intensities (and 95% confidence intervals) for dry season with no Influenza, vs wet season with Influenza .

Chapter 3

Models for burden estimation from time series counts

3.1 Introduction

This chapter contains a brief review of the time series methods to estimate burden of a disease (e.g. morbidity, mortality) attributable to a specific pathogen when dealing with count data. In fact, these methods are part of a broader class of models aimed at quantifying the association between two time series. Section 3.2 reviews the most important burden-estimation methods found in the literature, two of which are analysed in more depth in sections 3.2.1 and 3.2.2. Further, section 3.2.3 introduces the HHH modelling framework and its multivariate version, explaining how it addresses limitations of the previous two. Lastly, section 3.3 describes different model assessment approaches and section 3.4 introduces Granger causality.

3.2 Ecological studies

Ecological studies, i.e. observational studies analysing data at the population level rather than individual level, are often used to measure incidence and to identify drivers of a disease [177]. This choice is mainly made when individual-level data on the exposure are difficult to acquire, and population-wide routinely recorded measures can be taken as proxies for individual exposure. This is the case, for example, for studies on climate, that can be seen as a risk factor to which everyone is exposed equally and simultaneously.

With specific reference to respiratory infections, as anticipated in section 1.4, individual-level tests for the causing pathogen are not usually performed in LRTI patients, hence assessment of pathogen-specific burden relies on statistical inference. This analysis generally involves modelling time series of deaths or counts of syndromic healthcare contacts due to generic LRTIs with the aim of attributing a proportion of these cases to one or multiple pathogens.

However, LRTI incidence is characterised by strong seasonal patterns, with cyclic winter peaks in temperate areas of the world. An increased risk of LRTI is expected from a transient exposure to some viral or bacterial pathogen that exhibits the same seasonal variation. Hence, it is difficult to disentangle the natural LRTI burden, driven by meteorological factors, from that of viral and bacterial pathogens that also have a seasonal pattern of variation [100]. For example, as seasonal influenza viruses circulate every winter in temperate regions, the level of LRTI incidence in absence of flu is never observed. Such a quantity, sometimes called 'baseline', refers to the number of LRTI cases due to causes other than influenza. Temperature, humidity, pollution, sunshine hours and increased contact networks (e.g. schools) have all been identified as factors able to explain the annual fluctuations in respiratory disease incidence [46, 230], thus disentangling the contribution of each pathogen from a "seasonal confounding" adds to the challenge. A variety of regression methods have been proposed to estimate pathogen-specific excess morbidity and mortality, with a particular focus on the influenza virus [209, 104].

Beyond estimating pathogen-specific burden, a well-established problem, we consider these methods also with the aim of uncovering a possible interaction between pathogens using time series of population-level counts. Identification of co-infections at the individual level is unfeasible in routine clinical practice, due to the need for a time- and resource-consuming testing of respiratory specimens on all patients. However, for both viral and bacterial pathogens, aggregate information on positivity in the population can be retrieved from surveillance systems. Disentangling real interactions from spurious correlations due to shared driving variables is a common problem in ecology, where weak to moderate coupling is often observed in time series of species abundance. However, such an apparent synchrony might characterise also non-interacting species which share similar environments and are subject to forcing by environmental variables, e.g temperature and precipitation.

In the following sections we will review three categories of methods that have been used for this purpose in epidemiological settings, highlighting their advantages and limitations.

3.2.1 Cyclic regression models

In the general framework of burden estimation, population-level rates of non pathogen-specific outcomes are traditionally modelled by regression. Out-of-season outcome rates are used as a baseline for seasonal periods, so that the difference between observed rates and such baseline can be attributed to the circulating pathogen during a seasonal period [185, 194, 203].

The cyclic regression model was first introduced by Serfling [185]. Weekly counts Y_t are modelled as a function of calendar time t, including sine and cosine terms to represent seasonality, i.e.:

$$Y_t = \alpha + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{k}\right) + \beta_3 \cos\left(\frac{2\pi t}{k}\right) + \varepsilon_t, \tag{3.1}$$

α is the offset term for weekly person-time, k is the periodicity, generally selected so that the sine and cosine terms have a full period in one year, and ε_t is the error term, assumed to be normally distributed with constant variance. Coefficients α, β_1, ..., β_3 can then be conveniently estimated by ordinary least squares (OLS).

For example, the European Centre for Disease Control (ECDC) uses this method to estimate influenza excess mortality: the model is fitted to counts of deaths from weeks where influenza is not circulating, and such non-influenza mortality is then extrapolated to the influenza season [146]. The discrepancy occurring between the observed and estimated baseline during the excluded weeks is attributed to the influenza virus.

Several variations have been applied to this model, such as changes in the trigonometric functions to incorporate bi-annual periodicity [125] or the inclusion of a quadratic term for calendar time [193]. However, it is worth noting that the presence of trigonometric functions of time implies two assumptions: firstly, due to periodicity, the predicted outcome is forced to peak at the same time each year. Secondly, the winter increase in incidence must be equal in amplitude and duration to the summer decrease, relative to the yearly average, due to the functions' symmetry. To circumvent this limited flexibility, other factors that exhibit annual variation in intensity or timing, such as meteorological conditions or circulation of respiratory viruses, can be included in the model [103]. Explicitly adjusting for such factors

allows the strength and timing of incidence to vary across calendar years. Alternatively, indicator variables for month or season can be used in place of cyclic functions of time [184], even though such models may be overparameterised.

Beside the original limited flexibility, the Serfling model has a number of other weaknesses. First of all, information from some parts of the year, which could help better predict the winter baseline, is ignored in the estimation. To avoid the exclusion of such time periods, the *virological regression model* [27] is a generalisation of the Serfling model which includes information on the circulation of the pathogen of interest, z, as a covariate in equation (3.1):

$$Y_t = \alpha + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{k}\right) + \beta_3 \cos\left(\frac{2\pi t}{k}\right) + \beta_4 z_t + \varepsilon_t \tag{3.2}$$

Thanks to the presence of a proxy for pathogen circulation in the model, not only coefficients are now estimated using all the time points, but also z-related incidence can be directly inferred from the complete model, rather than extrapolated. Also, the 'baseline' can be obtained from the same model by setting $z = 0$.

Including a proxy for pathogen-specific circulation accounts for the presence and magnitude of that pathogen in the community more precisely than previous approaches. However, the validity of this strategy relies on the assumption that surveillance is consistent over time and adequately represents the true burden in the population [105]. If this assumption does not hold, apparent trends over time might be due to improved diagnostics or enhanced reporting rather than to a real spike in incidence. Finally, lagged effects of pathogen circulation can also be included, as a delay between infection and health-care contact is likely to occur, as symptoms are not developed immediately. However, this requires knowing the size of such time lag, or making a choice based on the model fit.

Thompson et al. [203] suggested that the outcome variable, weekly counts, could be more adequately modelled as Poisson distributed employing a log-link function, rather than using linear regression. The model is expressed as $Y_t \sim Poi(\mu_t)$ where

$$log(\mu_t) = log(pop_t) + \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{k}\right) + \beta_3 \cos\left(\frac{2\pi t}{k}\right) + \beta_4 z_t. \tag{3.3}$$

Such a link implies multiplicative effects of respiratory viruses on the outcome: the number of deaths increases exponentially with the number of laboratory reports, and the effects of different viruses and baseline are multiplicative on the number of deaths. As these assumptions are quite unrealistic, the plausibility of such a choice has been long debated. Gay

et al. [62] and Simonsen et al. [192] claimed that the model can only be linear and additive, since “total number of deaths is expressed as sum of the contributions from each pathogen, plus the seasonal background of deaths due to other causes". They suggested to use instead a generalized linear model (GLM) with a Poisson error distribution but identity link, so that the number of cases attributable to any virus are kept proportional to the laboratory reports. Multiple authors have then compared the estimates obtained with identity or log-link [204, 156], along with suggestions on use of splines to replace trigonometric functions and replace Poisson distribution with negative binomial to account for overdispersion [70, 33]. The model in (3.3) is routinely used to estimate the burden of influenza by the CDC [239] and elsewhere (e.g. [228]), nonetheless a discussion on the need of validation for such models is still ongoing [68, 236].

Finally, an important limitation of regression models is the assumption of independence among errors, unlikely to hold in this setting. When correlation in time series is not properly taken into account, variance is underestimated and confidence intervals are artificially narrower [164]. Hence, further precautions should be employed: seasonal block bootstrapping allows a correction of confidence intervals [172], whereas time-series methods are specifically tailored to deal with autocorrelation.

3.2.2 Time series methodology

Time series analysis includes a branch of statistical models that focus on temporal ordering of the observations and on their dependence over time. In fact, a time series is defined as a collection of observations made sequentially in time.

Let Y_t and Y_{t-h} be random variables describing the outcome of interest at time t and $t-h$ respectively. The quantity $h \in \{0,1,2,\dots\}$ is known as a *lag* [20], and the interest is often on the dependence between Y_t and Y_{t-h}. Such dependence can be quantified by the autocorrelation function $R_{t,t-h}$

$$R_{t,t-h} = \frac{E[(Y_t - \mu_t)(Y_{t-h} - \mu_{t-h})]}{\sigma_t \sigma_{t-h}} \tag{3.4}$$

where μ_t and σ_t represent the outcome mean and standard deviation at time t. If Y_t is stationary, i.e. μ and σ do not depend on time, the autocorrelation becomes only a function of the lag h. In other words, the autocorrelation for a given lag h will be the same for any t. A trend in the mean or a cyclical variation in time are the most common causes

of violation of stationarity. A visual diagnosis is facilitated by the correlogram, a plot of observed autocorrelation coefficients R_h against the lag h. In the presence of a trend R_h will not rapidly tend to zero as the lag increases, whereas it will show a sinusoidal pattern in case of seasonal data [20].

Since stationarity is a desirable property for most of the procedures used in time series analysis, preliminary ad-hoc transformations are often applied when non-stationary behaviour is detected, the most common methods being simple exponential smoothing [83] and the Holt-Winters procedure [93]. This is often needed when dealing with surveillance data, typically characterised by the replication of similar patterns across seasons. Once a stationary time series is obtained, the outcome at time t is generally regressed on its lagged values. An auto-regressive moving average (ARMA) model [12] is often employed for this purpose, and it was first used to model excess influenza mortality by Choi and Thacker [25]. The model takes the form

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \tag{3.5}$$

In the autoregressive component, the dependence of Y_t on the outcomes at the p previous time points is expressed through weights φ_p (usually decreasing going back in time). In the second part of the model, the regression error is replaced by a moving average term: the presence of a random disturbance at times $t-q$, persisting until time t, is expressed as a linear combination of current and past values of error variables $\{\varepsilon_t, \ldots, \varepsilon_{t-q}\}$, weighted by θ_q and assumed to be white noise (i.e. independent, identically distributed with zero mean).

An alternative approach through non-stationary time series is the autoregressive integrated moving average (ARIMA) model. This involves a two step procedure [20]. The dependent variable is first differenced to stabilize the mean, i.e. a new outcome variable is defined as $D_t = Y_t - Y_{t-d}$, where d represents the order of differencing. In a second step, the ARMA model described in equation (3.5) is estimated. ARIMA models are thus characterised by three parameters (p,d,q) identifying the order of autoregression, differencing and moving average terms respectively. Choice of the model order is made via an empirical procedure based on model fit. This makes ARIMA methods more suitable to the retrospective analysis of time series data rather than forecasting [213].

Beside the difficulty in model selection, inference for ARIMA models is not straightforward. Progressive transformations of the original data to achieve stationarity limit the interpretability of the coefficients, and maximum likelihood estimates are often obtained

assuming gaussianity of errors [20]. Finally, other important constraints include equally spaced data points and events occurring frequently, as the regression may perform poorly when trying to model data characterised by long gaps and sparsity [105].

ARIMA is a univariate technique, each time series being modelled separately. However, multiple series might relate to each other or to the same underlying process, and further methods are available to describe such relationships [143]. Given two time series Y_t and X_t, if the interest is on investigating the dependence of Y_t on X_t, we might simplistically include X_t as a covariate in model (3.5). The coefficient associated with X_t now expresses the effect of X_t on Y_t conditional on all the Y_{t-p}. More generally, ARIMAX models are an extension of ARIMA models allowing the inclusion of other time series as independent variables: Y_t is predicted from past lags of Y_t together with current and past lags of X_t [167]. They belong to the family of dynamic regression models.

If the two time series are stationary, the model takes the form

$$Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{i=0}^{b} v_i X_{t-i} + \sum_{i=0}^{q} \vartheta_i \varepsilon_{t-i},$$

where the noise terms ε_{t-i} are assumed to be independent from the input series X_{t-i}. If Y_t and X_t are not stationary, they are differenced before fitting the model with ARMA errors.

As many correlated covariates lead to collinearity problems, lags of X_t that usefully predict Y_t are preliminarily chosen. The cross-correlation function, formally defined as the correlation between X_{t-h} and Y_t for $h = \{0, 1, 2, \dots\}$, is usually employed for this purpose. For example, Gilca et al. [68] investigated the lag between influenza positivity and hospital admission, whereas Hendriks et al. [89] recently worked on ILI and IPD. Lagged variables best expressing the relation between the two time series are then included the regression model. For example, Hubert et al. [96] first estimated seasonal patterns of meningococcal counts, and in a second step regressed the detrended series on the relevant lags of ILI counts using an autoregressive error component. Similarly, Jackson et al. [104] added weekly counts of positive influenza tests as covariates after fitting an ARIMA model to the weekly incidence, nonetheless they found Serfling and virological regression models to be more accurate than the ARIMAX model when predicting the outcome.

3.2.3 Poisson branching processes

A third category of models derives from the theory of branching processes, through which dynamics of populations are modelled using generations as the time unit [78]. Let Y_t denote the size of the population at generation t, then a simple branching process can be written as

$$Y_t = \sum_{j=1}^{Y_{t-1}} Z_{j,t-1}$$

where $Z_{j,t-1}$ is the offspring of member j at generation $t-1$. The process Y_t only depends on Y_{t-1} and not on previous Y_{t-p}, hence it can be conveniently described through the conditional distribution $Y_t|Y_{t-1}$, with mean μ_t and variance σ_t^2. In particular, if the total population offspring Y_t is assumed to be Poisson distributed, Y_{t-1} acts directly on the conditional mean through an autoregressive parameter λ expressing the temporal dependence. In formulae,

$$Y_t|Y_{t-1} \sim \text{Poi}(\lambda Y_{t-1}). \tag{3.6}$$

Model (3.6) has been widely employed in modelling infectious disease data [55], since λ can be thought as an approximation for the average number of secondary infections generated by an infectious case in a totally susceptible population. Within branching processes theory, Held et al. [86] developed a new flexible modelling framework specifically for disease surveillance data, with the aim of retrospectively characterising epidemic evolution, in particular identifying the role played by covariates (e.g. seasonality), as well as prospectively forecasting outbreaks. They extended model (3.6) to a branching process with immigration component ν_t, i.e. $Y_t|Y_{t-1} \sim \text{Poi}(\mu_t)$ where

$$\mu_t = \nu_t + \lambda_t Y_{t-1}. \tag{3.7}$$

In equation (3.7), counts are modelled as the sum of two independent components: ν_t, endemic, and $\lambda_t Y_{t-1}$, epidemic. In the simplest case both ν_t and λ_t can be kept constant over time, however it might be of interest to further expand these terms. The average number of endemic cases ν_t can be parametrically modelled e.g. as a log-linear predictor that, multiplied by an offset such as population size pop_t, describes incidence due to regular trends and seasonal variations. Further, λ_t could also be time-varying, as the transmission of infection from time $t-1$ to time t might change over time, for example across seasons.

The decomposition of the contribution of several phenomena in additive components, along with the small number of parameters, makes interpretation very straightforward.

Compared to models presented in section 3.2.1, characterised by trigonometric functions, the presence of an autoregressive component in model (3.7) has the potential to better describe occasional outbreaks, as λ_t expresses the additional temporal dependence beyond the seasonality explained by ν_t [35, 85]. Moreover, the identity link allows preserving biologically-meaningful relationships among the quantities of interest.

Model (3.7) is further extended in Paul et al. [166] to analyse data from several pathogens. Denote by $Y_{i,t}$ the random variable representing the number of cases for pathogen i observed at weeks $t = 1,\ldots,T$. Let's assume for simplicity that there are only two pathogens, i.e. $i=\{1,2\}$. Paul et al. [166] assume that $Y_{i,t}|Y_{i,t-1} \sim Poi(\mu_{i,t})$ for both diseases, where each conditional mean is written as

$$\mu_{i,t} = \nu_{i,t} + \lambda_i Y_{i,t-1} + \tau_i Y_{j\neq i,t-1}. \tag{3.8}$$

The τ_i parameter expresses the association between the two time series and, if adjustments for seasonality are added in the endemic component, the interpretation of τ_i can more specifically refer to the association between the two pathogens after taking into account shared drivers. This setting can be generalised to incorporate more than two time series, estimating the association of the outcome of interest with more than one pathogen.

Furthermore, the Poisson distribution for the observation model can be replaced by a negative binomial for situations of overdispersed counts:

$$Y_t|Y_{t-1} \sim NegBin(\mu_t, \psi). \tag{3.9}$$

where ψ is the overdispersion parameter. Models (3.7) and (3.8) are both implemented in the R package *surveillance* [140] through the hhh4 function, named after the authors Held, Höhle and Hofmann, where maximum likelihood estimates are obtained via a (globally convergent) Newton-Raphson type algorithm.

More recently, Bracher and Held [13] suggested the inclusion of multiple lags for covariates, relaxing the assumption of temporal dependence limited to one week. Denoting by Q the number of lags considered, the mean is written as

$$\mu_t = pop_t \nu_t + \lambda \sum_{q=1}^{Q} w_q(y) Y_{t-q} + \tau \sum_{q=1}^{Q} w_q(x) X_{t-q} \tag{3.10}$$

where $w_q(y)$ and $w_q(x)$ are normalized lag weights defined according to a geometric structure, with parameters p_x and p_y shaping the respective exponential decays, i.e.

$$w_q(y) = \frac{p_y(1-p_y)^{q-1}}{\sum_{q=1}^{Q} p_y(1-p_y)^{q-1}} \tag{3.11}$$

Multivariate hhh model

The model in equation (3.8) can be also extended to deal with stratified time series: Meyer et al. [139] implemented a multivariate version for spatial disease spread. Let a be the group indicator, then two transmission components must be specified at this stage:

$$\mu_{t,a} = pop_{t,a}\nu_{t,a} + \lambda_a Y_{t-1,a} + \phi_a \sum_{k \neq a} c_{k,a} Y_{t-1,k \neq a} + \tau_a X_{t-1,a}. \tag{3.12}$$

In addition to the transmission of one pathogen within group a, quantified by λ_a, the transmission of the same pathogen across groups is explicitly incorporated: the coefficient ϕ_a, paired with the linear combination of disease cases in groups $k \neq a$ weighted by a factor $c_{k,a}$, represents the contribution of transmission from other population subgroups to disease in group a. Both transmission coefficients can be group-specific.

3.3 Predictive model assessment

In classical statistical theory, a probability distribution $p(\mathbf{Y} \mid \theta_0)$ is generally proposed to represent the process that generated the data $\mathbf{y}$. Maximum likelihood estimates $\hat{\theta}_0$ for the parameters are obtained in order to maximise the probability that data $\mathbf{y}$ have been observed under the assumed statistical model, nonetheless such a representation is almost never exact.

Model assessment involves measuring the discrepancy between the observed data, y_i for unit i, and the values fitted by the model, $g(y_i|\hat{\theta}_0)$ for some function $g()$, making use of a loss function d. Such a measure of distance between observed and fitted values can be summarised over all data points by $D(\mathbf{y}, g(\mathbf{y}|\hat{\theta}_0))$:

$$D(\mathbf{y}, g(\mathbf{y}|\hat{\theta}_0)) = \sum_{i=1}^{n} d(y_i, g(y_i|\hat{\theta}_0))$$

When considering a continuous random variable, a linear regression model is often employed to predict $\hat{y}_i$, and the difference $(y_i - \hat{y}_i)$ is called a residual. This quantity is

typically made positive through a quadratic loss function, and its sum over all data points i is called the residual sum of squares: $\text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$. Equivalently, the expectation of this loss can be computed, a quantity called the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

As a small RSS (or a small MSE) indicates a tight fit of the model to the data, this can be used as an optimality criterion in model assessment. More generally, when model parameters $\hat{\theta}_0$ have been estimated via maximum likelihood, the maximum value of the likelihood function $\widehat{L}_n(\hat{\theta}_0 \mid \mathbf{y}) = p(\mathbf{y} \mid \hat{\theta}_0)$ is itself a measure of goodness-of-fit, as we have derived $\hat{\theta}_0$ such that $\hat{\theta}_0 = \arg\max_{\theta \in \Theta} \widehat{L}_n(\theta\,; \mathbf{y})$. However, beyond assessing a model's absolute goodness-of-fit in relation to particular data, we also want to assess the performance of that model in relative terms with respect to other models (model comparison).

We usually assume that the underlying model generating the data belongs to a family of models $p(\mathbf{y}, \theta)$, and we are interested in selecting the formulation within that family which fits our set of observations $\mathbf{y}$ best. As any measure of goodness-of-fit is a monotonically increasing function of the number of parameters added to the model, an issue called overfitting, the statistics presented above cannot be used as a meaningful comparison of models with different numbers of independent variables.

In order to find a trade-off between the goodness-of-fit of the model and its simplicity, the model performance can be assessed on out-of-sample prediction instead. Such a process of measuring how well the model will generalize to an independent data set is called predictive assessment. Since new data are not often available, predictive assessment generally relies on cross-validation: if we call $\mathbf{y}^{\text{fit}}$ the portion of $\mathbf{y}$ used to fit the model, $\mathbf{y}^{\text{crit}}$ can denote the portion used for model criticism. Within the time series framework (leave-one-out) cross-validation translates into (one-step-ahead) forecasts, as the natural aim is finding the model that best predicts future outcomes based on the present and on the past.

Model performance is then measured in terms of the prediction errors $(y_i^{\text{crit}} - \hat{y}_i^{\text{crit}})$. Best predicting models have smaller prediction errors, i.e. higher agreement between predicted and observed values of $\mathbf{y}^{\text{crit}}$. For example, a model can be chosen as having the smallest the

out-of-sample MSE, sometimes called mean squared prediction error (MSPE):

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(y_i^{\text{crit}} - \hat{y}_i^{\text{crit}})^2.$$

Among statistical methods for the assessment of predictive performance, Akaike information criterion (AIC) [181] combines a measure of model fit, the maximum value of its likelihood function $\hat{L}$, with a measure of model complexity, the number of estimated parameters k:

$$AIC = 2k - 2log(\hat{L})$$

This is asymptotically equivalent to a cross-validatory loss function: a penalized loss function based on the deviance, with a penalty derived from a cross-validation argument.

Up to this point we have considered the model output to be a point estimate $g(\mathbf{y} \mid \hat{\theta}_0)$, generally the expected value of the outcome $\mathbf{Y}$. Other measures of goodness-of-fit are based instead on the the predictive distribution $p(\mathbf{y}, \hat{\theta}_0)$, which contains the complete set of probabilities associated to a number of different outcomes around the mean of the observed outcome [115].

Probabilistic modelling aims to maximise the sharpness of the predictive distribution while accounting for accuracy of the probabilistic predictions, i.e. the statistical consistency between the predictions and the observations. This procedure is often referred to as model calibration. Any loss function used to maximise calibration is then a function of the observed count y^{crit} and of the predictive distribution $Q = p(Y^{\text{crit}} \mid \hat{\theta}_0)$, and it is referred to as scoring rule $s(y^{\text{crit}}, Q)$. The model minimizing the expectation of $\text{E}_{\text{Y}^{\text{crit}}}\, s(y^{\text{crit}}, Q)$ will be closer to reporting the true probability for the prediction.

The logarithmic score $log(s(y_i^{\text{crit}}, Q)) = -logQ(y_i^{\text{crit}})$, i.e. the log-transformed predictive distribution evaluated at the observed outcome, is a standard choice in many situations, including when dealing with count data [40, 87]. It can also be written as cross entropy, a measure of divergence between the distribution of the observed outcome, p_i, and distribution of the predictions Q:

$$H(p_i, Q) = -\sum_{i}^{n} p_i \log Q_i.$$

3.4 Granger causality

All the methods presented in section 3.2 are based on cross-correlation, i.e. they quantify the magnitude of the linear dependence between two time series Y_t and X_{t-k}, for some lag $k = 0, \ldots, K$ possibly while taking into account the dependence from Y_{t-p} ($p \geq 1$) or from confounders Z_{t-k}. In 1969 Granger [75] suggested an alternative framework that uses predictability as opposed to correlation to identify relationships between time series variables. Variable X_{t-k} is said to "Granger cause" Y_t if the predictability of Y_t in some model declines when X_{t-k} is removed from the universe $U = \{X_{t-k}, Z_{t-k}, Y_{t-p}\}$ of all possible predictors. This implies that X_{t-k} contains unique information, i.e. not found in other variables, that can improve the prediction of Y_t.

In practice, Granger causality is usually tested based on linear regression models, even though extensions to nonlinear cases exist. An F-test is employed to compare the two models $p(Y_t \mid X_{t-k}, Z_{t-k}, Y_{t-p})$ and $p(Y_t \mid Z_{t-k}, Y_{t-p})$, testing the null hypothesis that X_{t-k}'s coefficients are not significantly different from zero. Granger causality can be inferred if the null hypothesis is rejected, i.e. X_{t-k} occurs before Y_t and the optimal model for Y_t improves when X_{t-k} is included. Such a definition strongly relies on the idea of directionality of time: any predictive factor can be a cause only if it occurs before the effect.

The key requirement of Granger causality is separability, namely that information about a causative factor is independently unique to that variable and can be removed by eliminating that variable from the model **??**. Such a requirement may not be met in dynamic systems with behaviors that are at least somewhat deterministic: if **X** is a cause for **Y**, information about **X** will be redundantly present in **Y** itself, which implies that the causal variable **X** cannot contain unique information, and that such information cannot be removed from U simply by eliminating **X**. A more detailed discussion of nonlinearity will be presented in chapter 8.

Moreover, the Granger-causality tests are designed to handle pairs of variables. Misleading results can be obtained when the true relationship involves three or more variables. For example, if both **X** and **Y** are driven by a common third process, we might get to a false positive conclusion of Granger causality, yet manipulation of one of the variables would not change the other. Similarly, in stochastic feedback systems it might result that Y_t Granger causes X_{t-k} but also X_t Granger causes X_{t-k}.

3.5 Conclusions

This chapter serves as a review of currently available methods aimed at estimating association between two or more time series, hence all the methods proposed and reviewed are applicable to surveillance data. When aiming to disentangle the contributions of different seasonal components, associations can be hard to identify in terms of linear correlations. Granger causality sets itself as a test for causal relationships beyond linear correlations.

We explore their application in the next chapter, investigating the interaction between viral and bacterial respiratory pathogens across age groups in England. This adds to existing evidence, that uses simpler models, and provides crucial evidence on age-specific pathogens interaction to inform stockpiling of antivirals and antibiotics.

Chapter 4

Estimating age-stratified influenza-associated IPD in England

4.1 Introduction

As introduced in sections 1.4 and 3.2, LRTIs are still responsible for a significant morbidity and mortality worldwide despite the availability of immunisation and antibiotics, and the contribution of each specific pathogen to LRTI burden must be estimated due to non-pathogen specific routine diagnostic practices. Counts of syndromic healthcare contacts for LRTI and pathogen-specific surveillance counts of detected infections can be combined for this scope via time series models: extensive efforts have been directed towards estimation of influenza burden [189, 39, 129].

However, especially in temperate countries of the world, widely studied respiratory pathogens such as Influenza virus, RSV, Rhinovirus and *Streptococcus Pneumoniae* are characterised by cyclical increases in autumn and winter months. Hence, investigation of the role of the interaction between pathogens relative to shared seasonality remains an open challenge [112, 230, 189]. We focus on the choice of a suitable methodological framework to address this question: Nicoli et al. [156] estimated the percentage of IPD cases attributable to influenza and RSV using a cyclic regression model, however the independence among observations they assume is unlikely to hold when modelling incidence of a transmissible pathogen. Hendriks et al. [89] proposed instead an ARIMA model, nevertheless preliminary transformations to the original counts limits interpretability of coefficients and the necessity of choosing model order via an empirical procedure based on model fit preclude ARIMA

methods as a sensible choice for our scope [213].

We proceed to investigate respiratory viral-bacterial interaction by using the flexible regression model introduced in section 3.2.3: weekly IPD counts are decomposed into an endemic component, with sine-cosine waves describing cyclic winter outbreaks, and an epidemic autoregressive component, where lagged IPD counts enter the model linearly using an identity link function. Time-varying covariates are also linearly added to the model, with the corresponding coefficients expressing the association between the outcome and each covariate after taking into account shared seasonality. We are interested in investigating the contribution of several pathogens to the incidence of IPD, as viruses other than influenza (RSV, rhinovirus) have been speculated to interact with *S. pnuemoniae* [195, 114]. Finally, as there is evidence that meteorological conditions such as temperature and humidity affect seasonality and intensity of outbreaks [187, 48], we replace sinusoidal functions with observed weather information.

As associations between pathogens have been suggested to be heterogeneous across age groups [151], we also consider the multivariate version of the modelling framework presented in 3.2.3. Beyond allowing estimation of age-specific associations, such a multivariate structure also permits decomposition of IPD transmission between and across age groups by incorporating contact patterns. In summary, compared to previous work [162, 190], we propose a phenomenological model that expresses IPD dynamics as a function of autoregressive components, viral infections, age-specific contact patterns and seasonal confounders without making strong assumptions on the transmission mechanism, aiming to provide a parsimonious characterisation of the drivers of IPD patterns over time.

4.2 Data

Influenza is generally diagnosed based on ILI, defined as the simultaneous presence of signs and symptoms such as high fever, cough and myalgia, however only virological testing allows the ascertainment of the responsible pathogen. For this reason, we estimated influenza incidence by combining two data sources. The RCGP RSC collects weekly numbers of GP consultations for several clinical diagnoses of communicable and respiratory diseases, including ILI. The population monitored by the RCGP RSC practices covers an average population of ≈ 1.4 million persons, 2.6% of England, considered to be representative of the national population in terms of age, gender, deprivation index and prescription patterns [34].

As part of routine virological surveillance, in general practices participating in the RCGP RSC scheme, a proportion of ILI cases is swabbed and the samples are tested for Influenza A (H1 or H3 subtypes), Influenza B, RSV and Human Metapneumovirus (hMPV) by the Public Health England (PHE) reference laboratory [34]. The number of specimens tested, and the number of positives for each virus, are stratified by week of test and age group to derive the proportion of virologically positive specimens. This proportion is then multiplied by ILI counts to compute the corresponding age and time specific consultations attributable to influenza.

S.pneumoniae (the pneumococcus) infection is often asymptomatic, as this is a commensal bacterium of the human nasopharynx, nonetheless its progression to the lower respiratory tract and blood can cause severe disease, namely IPD. In the UK, counts of positive isolates for a number of clinically significant pathogens are reported weekly to PHE by all the microbiology laboratories included in the national surveillance system, and stored in the Secondary Generation Surveillance System (SGSS) database. Counts of IPD, RSV and rhinovirus infections are extracted from SGSS. Consistency in testing over time and space is guaranteed by the “United Kingdom Standards for Microbiology Investigations", a diagnostic algorithm applied across laboratories to patients presenting with different clinical syndromes [51]. Finally, estimates of the population of England by age group, during each season, are obtained from the Office for National Statistics [161] while weather information such as daily Central England Temperature and daily England and Wales precipitation are downloaded from the MetOffice HadCET Data repository [165].

The time period considered ranges from 1st January 2009 to 31st December 2017, with the 2009 pandemic period defined to include the three waves, from week 15/2009 to week 26/2011 [174]. Disease incidence is categorised into five age groups: 0-4, 5-14, 15-44, 45-64 and 65+ years old, as in similar studies [156].

A total of 62,679 ILI consultations within the sentinel scheme and of 45,601 IPD cases nationwide have been notified over 9 years. The top panel of Figure 4.1 displays the temporal trend of all ILI and influenza-confirmed consultation rates respectively, where influenza-confirmed counts (referred to as “Flu" from now on) are obtained as just described. A clear seasonal pattern is visible, with regular outbreaks in the winter months and epidemics lasting 10-15 weeks, except for 2009 when the A/H1N1 pandemic started in spring. Virological testing is not systematically performed during the summer, hence the Flu data are quite sparse off-season. Nonetheless, it is evident how, even during winter, the influenza cases

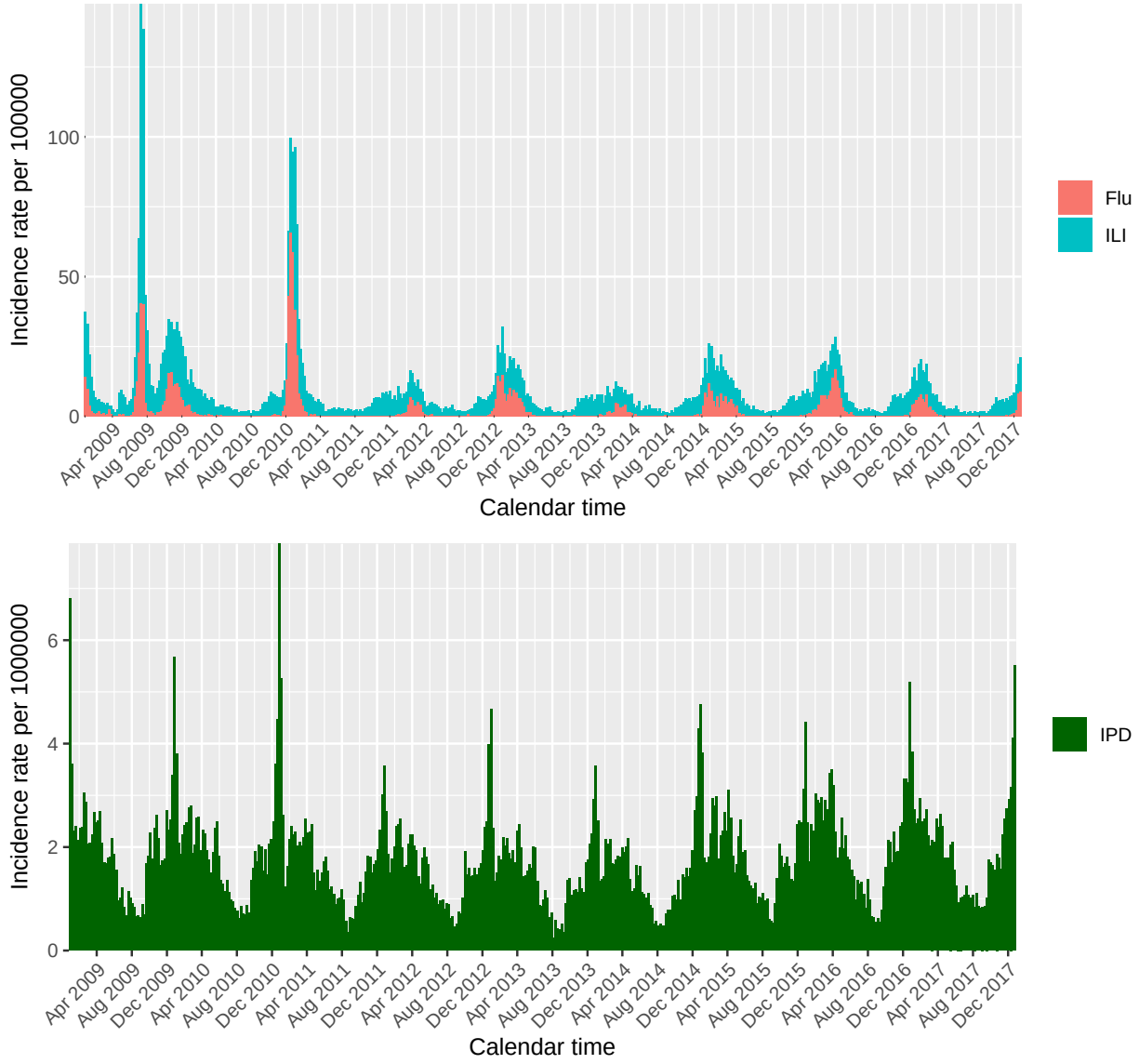


Fig. 4.1 ILI and Flu incidence rate in the top panel, IPD incidence rate in the bottom one.

do not closely mimic the ILI curve, confirming the non-specificity of the ILI diagnosis. In the IPD time series (Figure 4.1, bottom panel), peaks appear to be similar across seasons both in terms of amplitude and timing, with a gradual increase of cases from autumn to a winter peak, followed by a decline in summer. The incidence rate per 1,000,000 population is plotted in this case, as IPD is rare. The observed time series for RSV and rhinovirus are plotted in appendix A, Figure A.2.

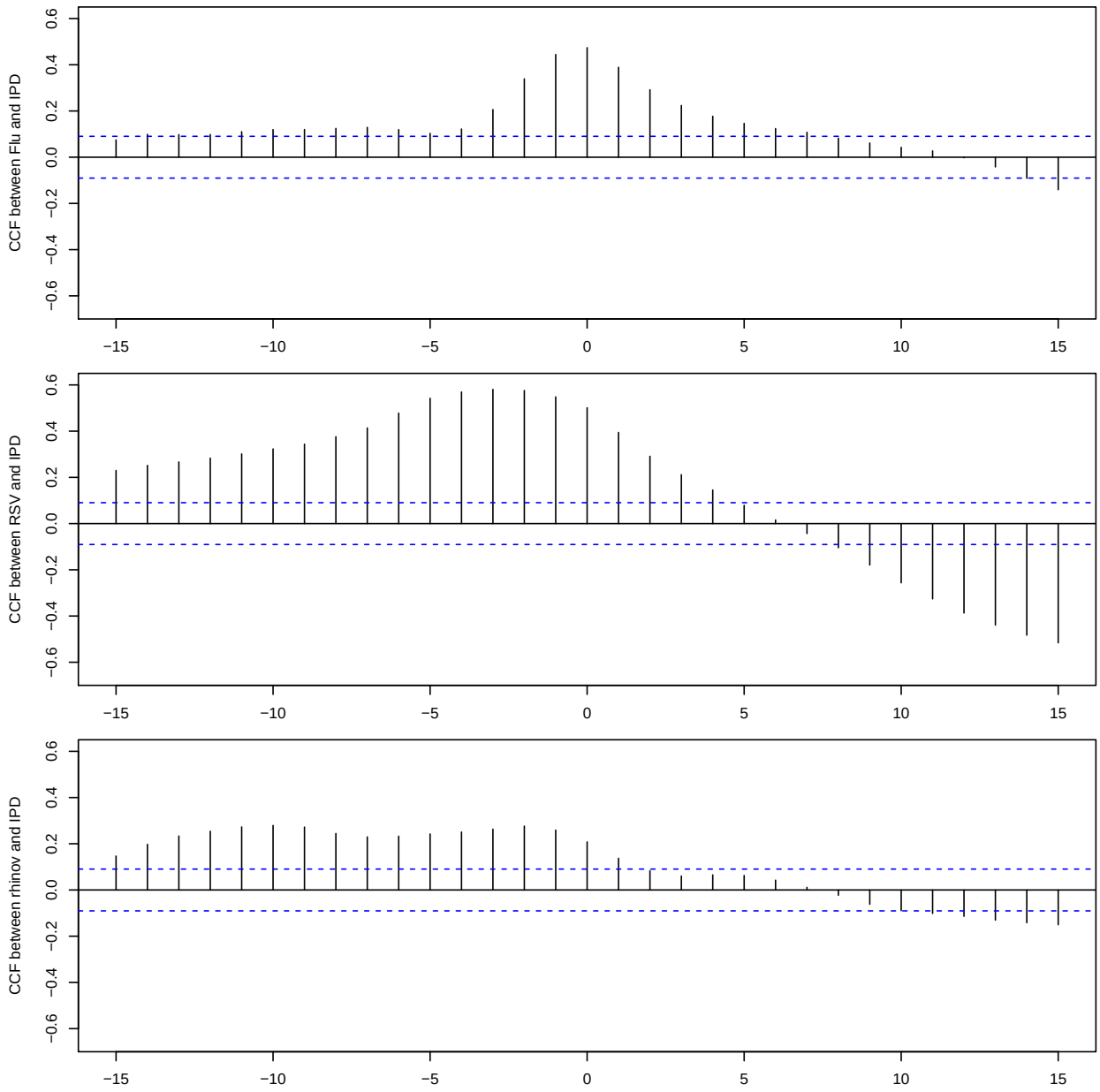


Fig. 4.2 Cross-correlation between each virus and IPD for up to 15 weeks of lag. Top panel: flu and IPD. Middle panel: RSV and IPD. Bottom panel: rhinovirus and IPD.

4.2.1 Bivariate time series analysis

Cross-correlations between each virus and IPD are plotted in the three panels of Figure 4.2: a strong correlation between flu and IPD with no clear lag is detected, confirming the overlapping trends of Figure 4.1. Correlation with RSV is equally strong but for negative lags, while correlation with rhinovirus is smaller, yet more marked for negative lags. Cross-correlation

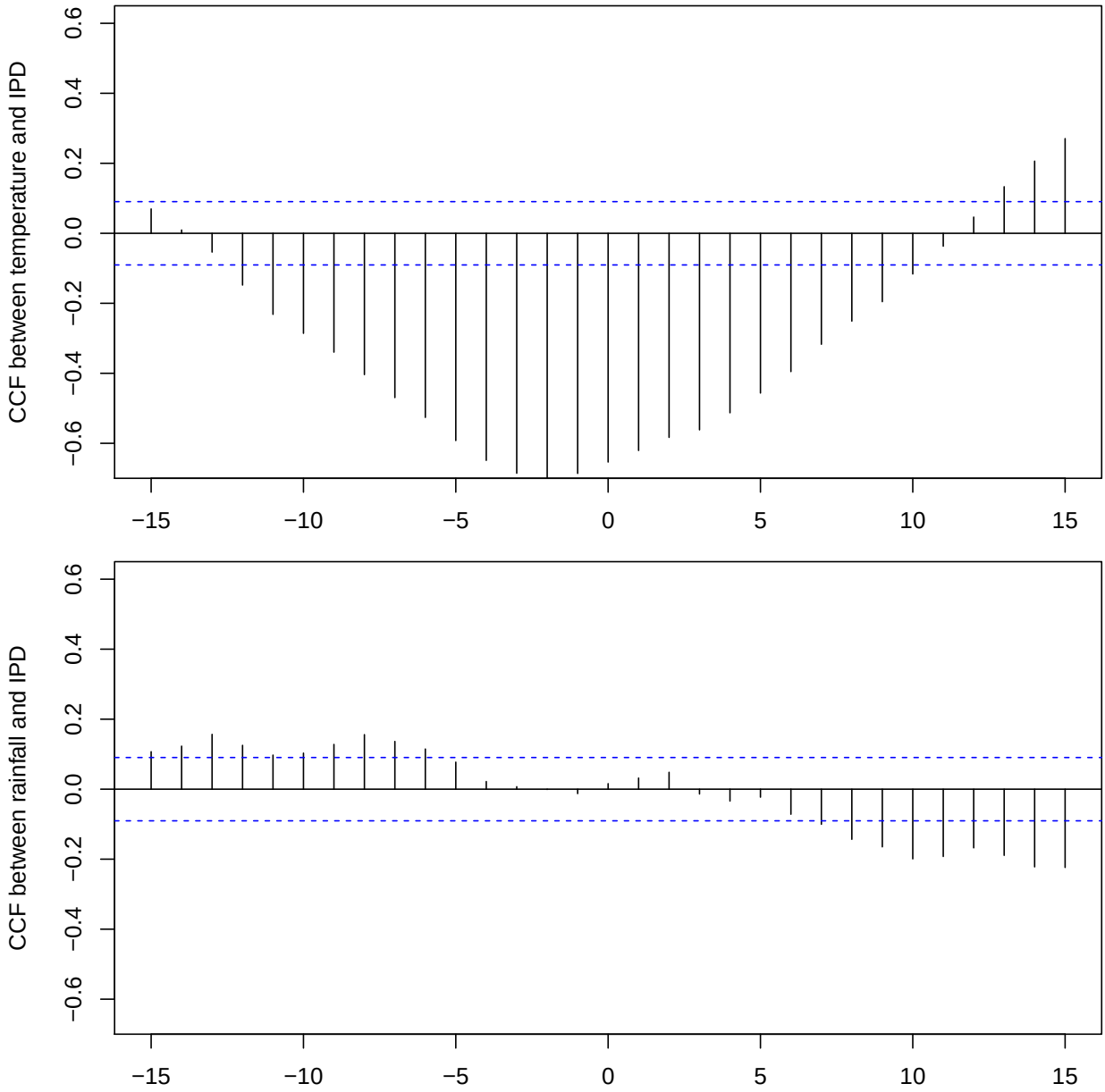


Fig. 4.3 Cross-correlation between weather variables and IPD for up to 15 weeks of lag.

between weather variables and IPD is plotted in Figure 4.3: a strong negative correlation is detected for temperature in correspondence of negative lags, whereas correlation is barely significant for rainfall. We then test for the presence of Granger causality, i.e. compare the model in which IPD is explained by k lags of IPD only, with the same model to which k lags of one explanatory variable are added. P-values of each test are summarised in Table 4.1: in the top half of the table we see that including 1 to 5 lags of each variable, except for rainfall,

improves model fit. On the bottom half of the table we test whether the opposite is true: we model each explanatory variable alone, and test whether IPD lagged counts improve model fit. The significant results for RSV and rhinovirus (feedback system) suggest that, along with IPD, they might be driven by a common process, hence this Granger causality might be a false positive. On the other hand, the relationship for flu and temperature is unidirectional towards IPD, suggesting any change in their intensity might lead to changes in IPD.

k	flu	RSV	rhinovirus	temp	rain
1	0.0035	0.0000	0.0001	0.0000	0.8062
2	0.0000	0.0000	0.0022	0.0000	0.8821
3	0.0000	0.0000	0.0002	0.0000	0.9226
4	0.0000	0.0000	0.0008	0.0000	0.9357
5	0.0000	0.0000	0.0027	0.0000	0.4222
1	0.5323	0.0000	0.0000	0.0043	0.4264
2	0.4217	0.0007	0.0003	0.2202	0.5959
3	0.5024	0.0009	0.0001	0.5125	0.2220
4	0.6966	0.0033	0.0000	0.1665	0.1556
5	0.7886	0.0030	0.0000	0.1739	0.2604

Table 4.1 P-values from a Granger causality test assessing the predictability of IPD as a function the listed covariates

4.3 Analysis strategy

We propose the **hhh** approach presented in section 3.2.3 to model IPD and Flu counts. We start by considering the time series referring to the entire population, without age stratification, and we model $IPD_t|IPD_{t-1} \sim Poi(\mu_t)$ where

$$\mu_t = pop_t \nu_t + \lambda Y_{t-1} + \tau Flu_{t-1} \tag{4.1}$$

Alternatively, $IPD_t|IPD_{t-1} \sim NegBin(\mu_t, \psi)$. We test the usefulness of an overdispersion parameter by assessing which model formulation fits our data best in terms of AIC over all the study period. Then, considering the mean is decomposed as in equation 4.1, we compare different expressions for the endemic component ν_t. Yearly seasonality for weekly data is described through the general formulation

$$log(\nu_t) = \alpha + \sum_{s=1}^{S} \gamma_s sin\Big(\frac{2\pi st}{52}\Big) + \delta_s cos\Big(\frac{2\pi st}{52}\Big) \tag{4.2}$$

where α is an intercept and γ_s and δ_s quantify the amplitude of the sine-cosine waves. We first assess the optimal number S of trigonometric functions to be included, and in a second stage we check whether replacing the sine-cosine waves with rainfall and temperature information better characterises differences across winters. We then consider the inclusion of multiple lags both for Flu and IPD, following the formulation in equation 3.10: lags $q = 1, \ldots, Q$ where $Q = 5$ are assessed, i.e. including incidence of Flu and IPD up to 5 weeks before time t. Finally, the additional contributions of RSV and rhinovirus are considered by sequentially adding the number of detected infections to the selected model for IPD and influenza.

We also compare the accuracy of the different model formulations in terms of one-step-ahead forecasts using the logarithmic score, as described in section 3.3. We select 30 weeks as the initial time window of data used as training period, and for each training set of length $j = 30, \ldots, 440$ we refit the model, produce a one-step-ahead forecast and compute the logarithmic score $log(s_j(p,x))$. The best model is chosen by taking the expected value over j.

We then model IPD counts in age group a, $IPD_{t,a}$, where $a \in$ {0-4, 5-14, 15-44, 45-64, 65+}, using the multivariate version of the hhh model. For consistency, we use for all age groups the distributional assumption and the endemic component that fitted the univariate time series best, and the mean component is now formulated as

$$\mu_{t,a} = pop_{t,a}\nu_{t,a} + \lambda_a IPD_{t-1,a} + \phi_a \sum_{k \neq a} c_{k,a} IPD_{t-1,k \neq a} + \tau_a Flu_{t-1,a}. \tag{4.3}$$

In order to account for heterogeneity of contact patterns, counts of disease in groups $k \neq a$ are weighted by the element $c_{k,a}$ of the POLYMOD contact matrix, a measure of social distancing between group k and a [149]. We initially assume transmission coefficients to be age-specific as, despite accounting for contact patterns, some age groups are known to be more susceptible to infection than others. Similarly, heterogeneity across groups can be allowed for the remaining model parameters, as the interaction between influenza and *S.pneumoniae* has also been suggested to vary with age [43]. We then use model selection via AIC and $log(s(P,x))$ to evaluate whether some coefficients could be shared across age groups: model fit is assessed in a sequential way, testing at each stage which of the components leads to a larger AIC reduction when associated with non age-specific coefficients.

The model in equation (4.3) is simultaneously fitted to the five age groups, and code was not available for the situation where multiple covariates are added. We developed our

own algorithm, simultaneously fitting models for different strata incorporating the contact structure. Similarly to the hhh4 function, we also obtained maximum likelihood estimates via a (globally convergent) Newton-Raphson type algorithm. To ensure positivity, parameters are optimized on the log-scale, i.e. $log(\psi)$ and $log(\lambda)$ are used. Uncertainty about the proportions of IPD cases attributable to each virus is estimated by resampling n=10,000 datasets from the fitted model and taking the 95% confidence intervals (CIs) to be the empirical 2.5% and 97.5% percentiles across the resampled datasets.

4.4 Results

4.4.1 Model choice: endemic waves and lagged covariates

A summary of model comparison is presented in Table 4.2: starting from a Poisson distributional assumption and one set of trigonometric functions (model A), we first add an overdispersion parameter (model B); more complicated versions of the endemic component are then assessed by replacing trigonometric waves with weather variables (model C) and including multiple lags for them (model D). Great improvement in model fit is obtained by allowing overdispersion, whereas we see no gain in adding either Flu or IPD lagged counts when the lag $q > 1$. Model fit is improved instead when adding lagged values for rainfall and temperature, however the parameter representing the decline in weight attributed to lagged values is optimally chosen to be $p_{weather} = 0.8$, suggesting that only 20% of the weight is attributed to observations more than one week before. Evaluating the model in terms of one-step-ahead forecasts, we also find mean $log(s(P,x))$ to be minimum for the endemic formulation including weather information, with lags weighted according to $p_{weather} = 0.8$ (model D).

	distr	endemic	covar	AIC	log(s(P,x))
A	Poi	S=1	Flu	5107.61	5.805
B	NB	S=1	Flu	4043.95	4.408
C	NB	rain+temp, lag=1	Flu	4029.19	4.400
D	NB	rain+temp, lags=5 ($p_{weather}$=0.8)	Flu	4027.82	4.390
E	NB	rain+temp, lags=5 ($p_{weather}$=0.8)	Flu+rhinov	3997.93	4.361
F	NB	rain+temp, lags=5 ($p_{weather}$=0.8)	Flu+rhinov+RSV	3992.95	4.334

Table 4.2 Model comparison in terms of AIC and one-step ahead forecast (log(s(P,x)))

4.4.2 Estimated influenza impact

Fitted values for all components according to model formulations B and D are shown in Figures 4.4 and 4.5, whereas predictive distributions for one-step-ahead forecasts of model D are presented in Figure 4.6. The number of IPD cases attributed to Flu during the entire study period is as low as 199 according to model D including weather variables, i.e. 0.45% (CI <0.01%-1.59%) of all the IPD cases. However, 100 of these cases happened during the three pandemic waves, 0.83%, CI <0.01%-2.94%, of all the observed IPD cases in that period, suggesting that the pandemic strain might have been responsible for an increased incidence. As a sensitivity analysis, we select Flu counts referring only to the three pandemic waves: the increase in AIC is minimal compared to model D including Flu counts over all the study period, suggesting that the role of seasonal Flu is marginal. We also consider each season as a separate covariate, with results plotted in Appendix A, Figure A.1.

4.4.3 Rhinovirus and RSV

Finally, we investigate whether other viruses also interact with *S.pneumoniae*: the number of rhinovirus (model E in Table 4.2) and RSV (model F) infections are sequentially added to the selected model D. Rhinovirus alone greatly enhances the fit to the data, and the inclusion of RSV on top of Flu and rhinovirus still results in model improvement. Hence, the best fitting model (F) for mean IPD counts at time t takes the form

$$\begin{aligned}\mu_{IPD,t} = pop_t \Big[exp(\alpha + \gamma \sum_{q=1}^{5} w_q(weather) temp_{t-q} + \delta \sum_{q=1}^{5} w_q(weather) rain_{t-q}) \Big] + \\ + \lambda IPD_{t-1} + \tau Flu_{t-1} + \theta rhinovirus_{t-1} + \zeta RSV_{t-1}\end{aligned} \tag{4.4}$$

with overdispersion parameter ψ and decay parameter for $w_q(weather)$ fixed to $p_{weather}$=0.8. Point estimates and standard errors for the coefficients are reported in Table 4.3, while relative contributions are pictured in Figure 4.7: rhinovirus explains 6.97% (CI 4.27%-10.28%) of all the IPD cases, 2.48% (CI 0.51%-4.52%) are attributed to RSV and only 0.67% (CI <0.01%-1.69%) to Flu. Overall, the three viruses account for 10.12% (CI 7.18%-13.77%) of IPD cases at population level.

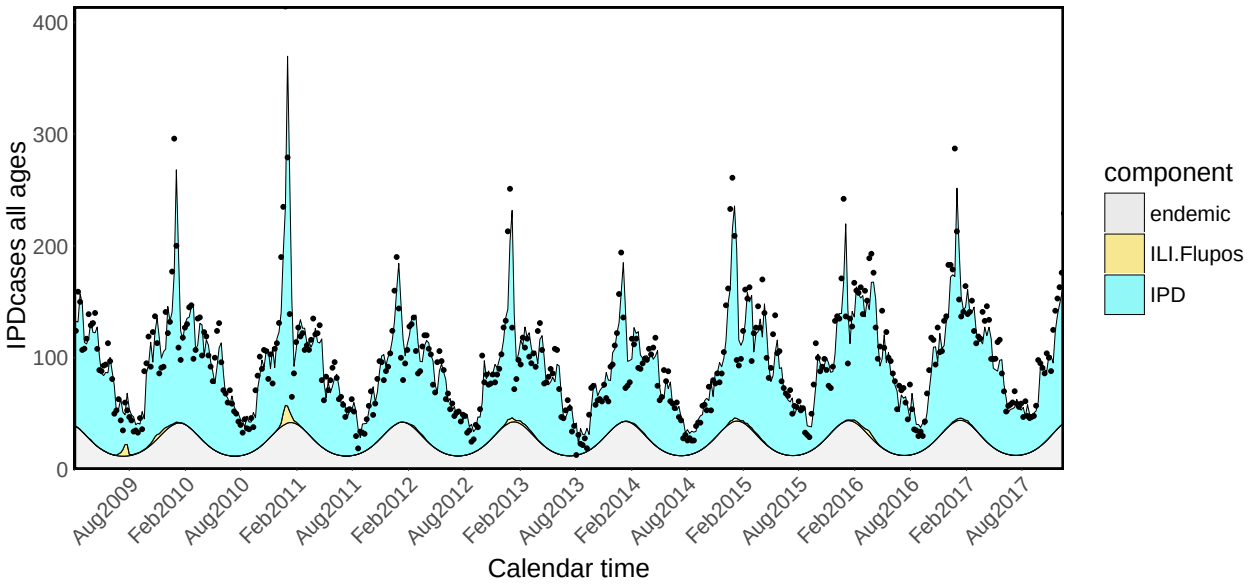


Fig. 4.4 Model (B) of IPD and Influenza with one set of trigonometric functions

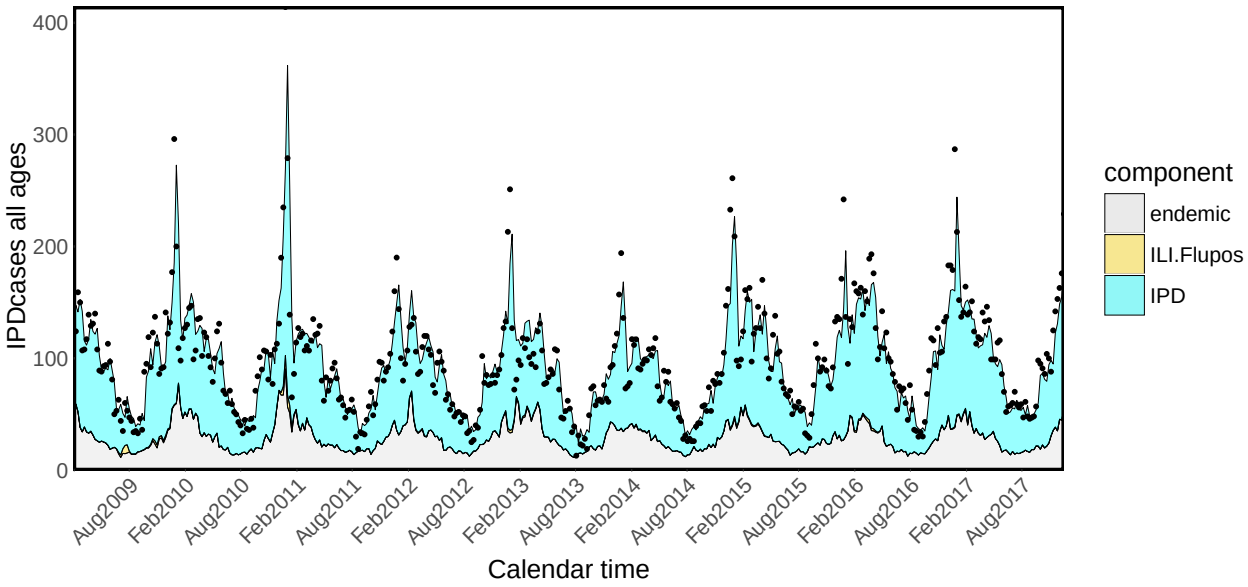


Fig. 4.5 Model (D) of IPD and Influenza with rainfall and temperature

4.4.4 Age-specific analysis

Selected plots displaying age-specific incidence can be found in Appendix A, Figures A.3-A.5. For consistency, we use for all age groups the distributional assumption and the endemic component that fitted the univariate time series best (model D). Thus, when considering attribution of IPD to Flu, model selection starts by considering the model in equation (4.3):

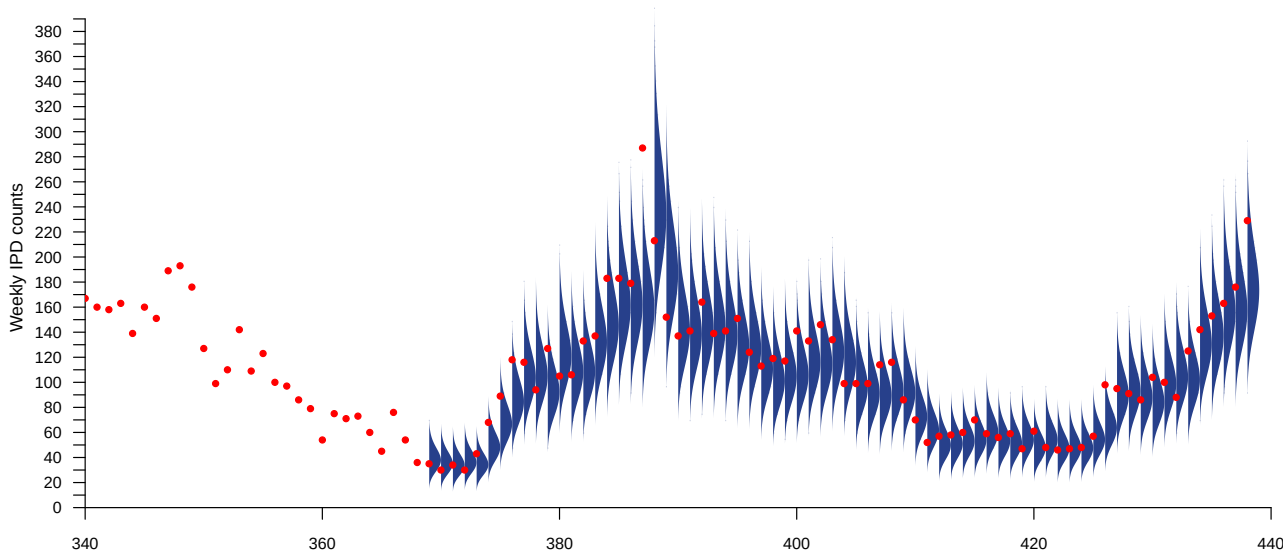


Fig. 4.6 One-step-ahead predictive distribution and observed values

	Estimate	Std. Error
α	-3.1831	0.0066
γ	-0.4572	0.0052
δ	-0.0838	0.0034
$log(\psi)$	3.5064	0.0080
$log(\tau)$	-1.2946	0.3240
$log(\theta)$	2.9821	0.0193
$log(\zeta)$	1.8944	0.0742
$log(\lambda)$	5.8521	2e-04

Table 4.3 Coefficient estimates for model (I), including Flu, rhinovirus and RSV as covariates.

$IPD_{t,a}|IPD_{t-1,a} \sim NegBin(\mu_{IPD,t,a}, \psi_a)$ where

$$\mu_{IPD,t,a} = pop_{t,a}\left[exp\left(\alpha_a + \gamma_a \sum_{q=1}^{5} w_q(weather)temp_{t-q} + \delta_a \sum_{q=1}^{5} w_q(weather)rain_{t-q}\right)\right] + \\ + \tau_a FluA_{t-1,a} + \lambda_a IPD_{t-1,a} + \phi_a \sum_{k \neq a} c_{k,a} IPD_{t-1,k \neq a} \tag{4.5}$$

However, this requires estimating 35 coefficients, not a very parsimonious option. Hence, we try model reduction by testing whether any of the coefficients could be the same across groups. Full model comparison is reported in Table 4.4. AIC decreases from 13218.85 (model G, with all age-specific coefficients) to 13216.32 by using a shared rainfall coefficient for any age, i.e. $\delta_a = \delta$ (model H). Finally, the utility of multiple lags for Flu and IPD is con-

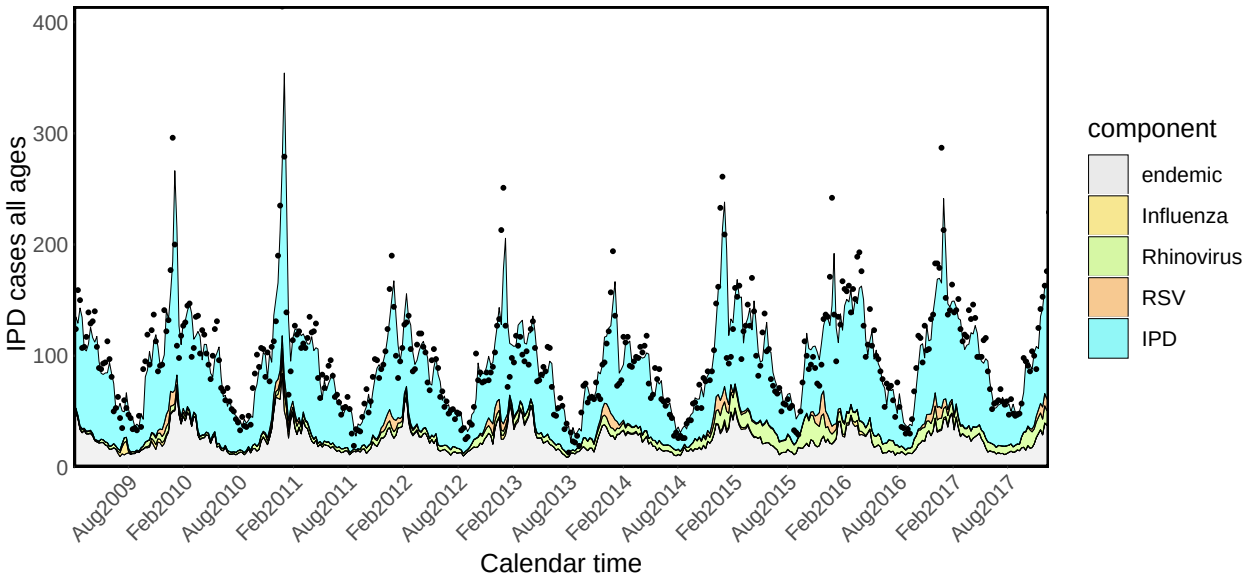


Fig. 4.7 Model (F) including Influenza, rhinovirus and RSV

sidered, but once again a benefit from including past values only pertains to weather variables.

	covar	shared pars	AIC
G	Flu	all age-spec	13218.85
H	Flu	$\delta_a = \delta$	13216.32
I	Flu	$\delta_a = \delta$, $\tau_{<5} = \tau_{65+} = 0$	13212.32
J	Flu+rhinov	$\delta_a = \delta$, $\tau_{<5} = \tau_{65+} = \theta_{5-14} = \theta_{15-44} = 0$	13160.70
K	Flu+rhinov+RSV	$\delta_a = \delta$, $\tau_{<5} = \tau_{65+} = \theta_{5-14} = \theta_{15-44} =$ $= \zeta_{5-14} = \zeta_{15-44} = 0$	13143.67

Table 4.4 Multivariate model comparison in terms of AIC

Estimated coefficients and standard errors for model H are shown in Tables 4.5 and 4.6. The τ_a parameters associated with influenza are quite heterogeneous across age groups, showing an inverse-U shaped tendency: almost null in young children and the elderly, and more prominent in other age groups. However, due to the very small size and associated large uncertainty of the parameters $\tau_{<5}$ and τ_{65+}, we refit the model fixing them to zero (model I). The attributed proportions of IPD cases estimated from this model are reported in Table 4.7, estimated coefficients and standard errors are shown in Tables 4.8 and 4.9, fitted values for all age groups are plotted in Figures 4.8-4.10 and predictive distributions in the Appendix, Figures A.6 and A.7.

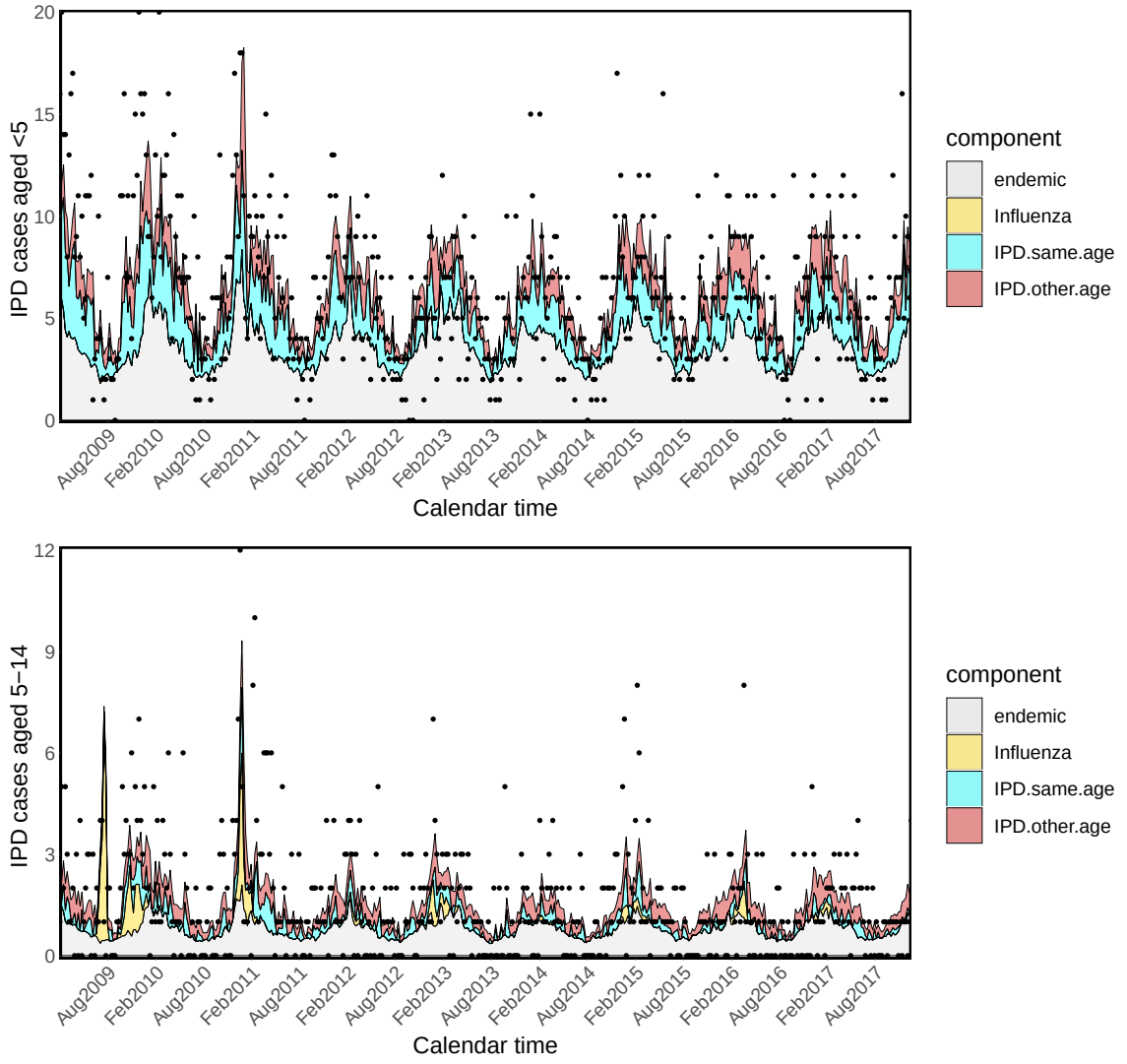


Fig. 4.8 Model I: Fitted IPD values for infants and school-age children

Importantly, according to model I, IPD is driven by Flu in school-age children (8.40%, CI 4.12%-13.66%) and adults aged 15-44 (3.55%, CI 1.64%-5.76%), and these components are strikingly higher in the pandemic period: 18.30% (CI 9.43%-28.16%) and 6.07% (CI 2.83%-9.76%) respectively.

Adding rhinovirus in the best fitting model I leads to the biggest AIC reduction, from 13216.32 to 13167.31, when its contribution is quantified by an age-specific coefficient θ_a. Lastly, the addition of RSV further contributes to AIC reduction (13153.89). Hence, the final

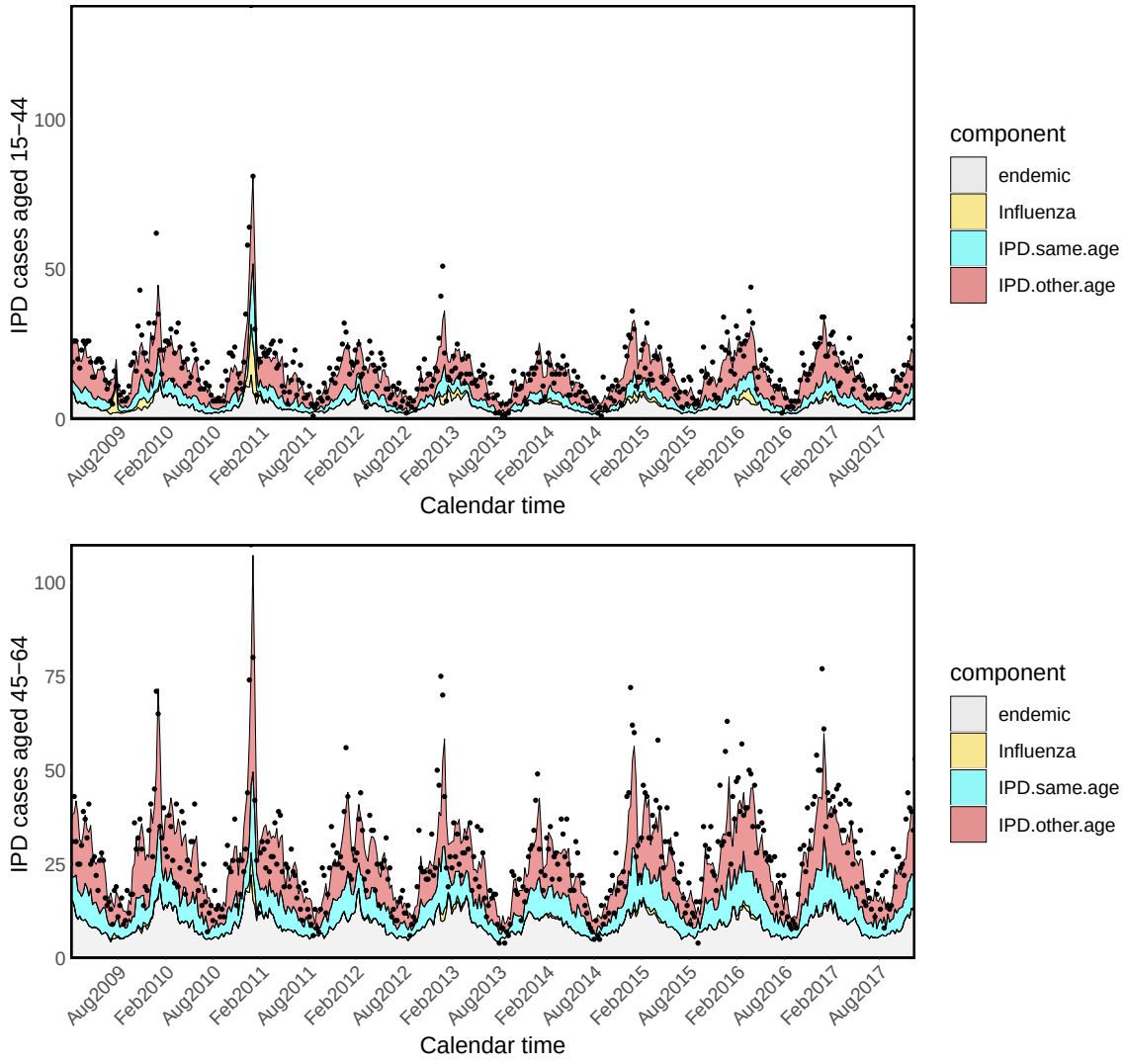


Fig. 4.9 Model I: Fitted IPD values for the 15-44 and 45-64 age groups

model takes the form

$$\mu_{IPD,t,a} = pop_t \nu_{t,a} + \tau_a Flu_{t-1} + \theta_a rhinovirus_{t-1} + \zeta_a RSV_{t-1} + + \\ + \lambda_a IPD_{t-1} + \phi_a \sum_{k \neq a} c_{j,i} IPD_{k \neq a,t-1} \tag{4.6}$$

where $\nu_{t,a} = exp\left[\alpha_a + \gamma_a \sum_{q=1}^{5} w_q(weather) temp_{t-q} + \delta \sum_{q=1}^{5} w_q(weather) rain_{t-q})\right]$. As for the model with only Flu, due to large uncertainty about coefficients close to 0, the coefficients θ_{5-14}, θ_{15-44}, ζ_{5-14} and ζ_{15-44} are fixed to zero (models J and K). Fitted values for all age groups are plotted in Appendix A, Figures A.8-A.12, coefficients and standard errors are listed in Tables A.1 and A.2 while the relative contribution of the components is de-

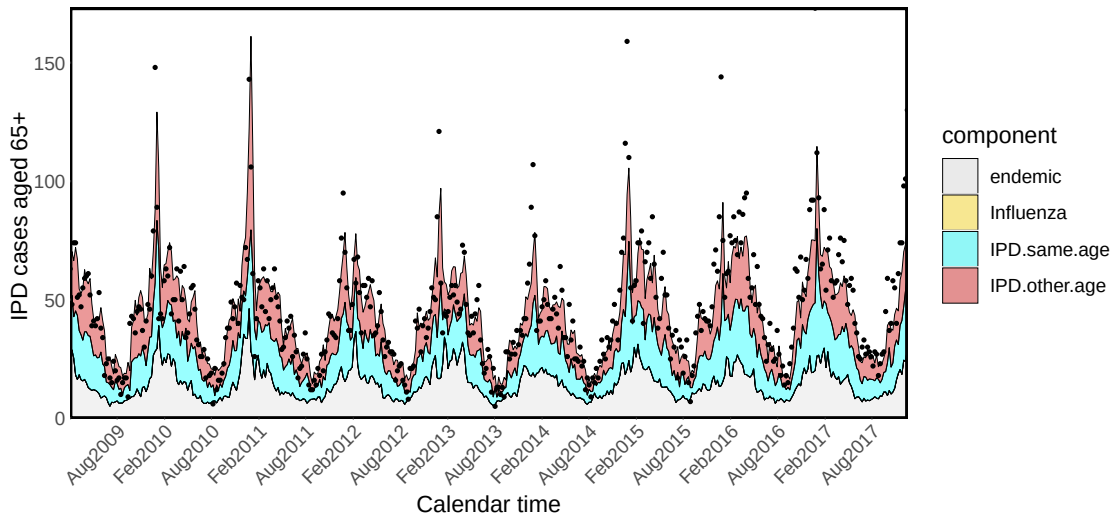


Fig. 4.10 Model I: Fitted IPD values for the elderly

Age	α	γ	δ	$log(\psi)$	$log(\tau)$	$log(\lambda)$	$log(\phi)$
<5	-2.282	-0.317	-0.037	2.513	-13.847	2.182	1.291
$5-14$	-4.422	-0.369	-0.037	1.590	-3.262	2.283	1.364
$15-44$	-4.035	-0.477	-0.037	3.024	-1.629	3.729	4.093
$45-64$	-2.776	-0.342	-0.037	3.173	-2.150	3.479	3.610
$65+$	-1.938	-0.464	-0.037	3.188	-6.714	3.423	4.246

Table 4.5 Model I: Coefficient estimates for the age-specific model of IPD including Flu. Since Flu coefficients $\tau_{<5}$ and τ_{65+} were very small, we refit the model fixing them to 0, to make sure the other parameter estimates are not sensitive to such an assumption.

scribed in Table 4.10. Model K shows that the association between RSV and IPD is strongest in the elderly (3.91%, CI 1.83%-6.38%, of cases in 65+ and 4.18%, CI 1.58%-6.91% of cases in 45-64), and rhinovirus plays an important role in the same age groups: 5.43% (CI 2.23%-8.91%) in 45-64 and 5.68% (CI 3.03%-8.32%) in 65+.

Age	α	γ	δ	$log(\psi)$	$log(\tau)$	$log(\lambda)$	$log(\phi)$
<5	0.126	0.040	0.024	0.197	177.157	0.179	0.398
$5-14$	0.254	0.086	0.024	0.269	0.354	0.301	0.482
$15-44$	0.182	0.063	0.024	0.148	0.307	0.264	0.127
$45-64$	0.119	0.039	0.024	0.126	1.048	0.221	0.141
65+	0.111	0.034	0.024	0.101	-	0.158	0.147

Table 4.6 Model I: Standard error estimates for the age-specific model of IPD including Flu. Uncertainty around coefficients $\tau_{<5}$ and τ_{65+} was not well estimated.

Age	Pnc transm		Influenza A	
	within group	across groups	overall	AH1N1pmd09
<5	26.42 (16.16 - 34.49)	18.73 (4.67 - 33.14)	0.00	0.00
$5-14$	15.70 (5.32 - 24.17)	27.32 (2.18 - 55.97)	8.40 (4.12 - 13.66)	18.30 (9.43 - 28.16)
$15-44$	19.47 (10.33 - 27.07)	50.67 (38.79 - 63.16)	3.55 (1.64 - 5.76)	6.07 (2.83 - 9.76)
$45-64$	23.65 (14.79 - 31.03)	41.64 (31.61 - 51.49)	0.92 (<0.01 - 2.94)	1.19 (<0.01 - 3.78)
65+	33.02 (24.89 - 39.88)	34.45 (26.02 - 43.24)	0.00	0.00

Table 4.7 Model I: Relative proportions (%) of IPD cases attributed to pneumococcal transmission within and across age groups, and to influenza overall or in the pandemic period

4.5 Discussion

Using English surveillance data, we quantify the magnitude of the interaction between influenza virus and *S.Pnuemoniae* in seasonal and pandemic settings by proposing a multivariate extension of the **hhh** modelling framework. Such interaction is estimated to be quite small when looking at population-wide counts (model D). These results are consistent with previous research, showing a small association at aggregate level [112]. Interestingly, we find evidence to support the hypothesis of an age-specific interaction [151], the contribution of Flu towards IPD being significant in school-age children and adults aged 15-44 but not in other age groups (model I). Moreover, the components of IPD explained by influenza are strikingly higher during the 2009 pandemic period in the same age groups. This supports the findings of Weinberger et al. [232]. Other viruses also appear to interact with *S.pneumoniae* with various intensities across age groups: both RSV and rhinovirus play an important role in 45-64 and 65+ year-olds (models F and K respectively). Such findings support previous evidence of interplay among these pathogens, with differential behaviour across ages [109, 231].

Compared to Serfling-type models, an important advantage of the modelling framework used here is the potential to account for correlation between observations thanks to the inclusion of an autoregressive term. The importance of such a component, which represents

Age	α	γ	δ	$log(\psi)$	$log(\tau)$	$log(\lambda)$	$log(\phi)$
<5	-2.2818	-0.3173	-0.0372	2.5131	-	2.1816	1.2908
5-14	-4.4221	-0.3688	-0.0372	1.5905	-3.2617	2.2835	1.3644
15-44	-4.0347	-0.4770	-0.0372	3.024	-1.6285	3.7285	4.0934
45-64	-2.7756	-0.3419	-0.0372	3.1729	-2.1502	3.4786	3.6104
65+	-1.9382	-0.4640	-0.0372	3.1881	-	3.4232	4.2462

Table 4.8 Model I: Coefficient estimates for the age-specific model of IPD including Flu

Age	α	γ	δ	$log(\psi)$	$log(\tau)$	$log(\lambda)$	$log(\phi)$
< 5	0.0033	0.0017	0.0012	0.0385	-	0.0109	0.0142
$5-14$	0.0145	0.0092	0.0012	0.0718	0.1035	0.0517	0.0222
$15-44$	0.0063	0.006	0.0012	0.0220	0.0661	0.0068	9e-04
$45-64$	0.0128	0.0113	0.0012	0.0158	0.7316	0.0036	0.001
$65+$	0.0017	0.0011	0.0012	0.0101	-	0.0018	0.0012

Table 4.9 Model I: Coefficient standard errors for the age-specific model of IPD including Flu

pneumococcal disease transmission, is undoubted: our findings suggest that 50.70% (CI 38.19%-63.20%) of pneumococcal disease in adults aged 15-44, potential parents of young children, is transmitted from other age groups. Transmission within group, on the other hand, prevails in pre-school children and 65+ year-olds: 26.32% (CI 16.24%-33.95%) and 23.75% (CI 14.97%-30.68%) respectively (model K). We speculate this could be due to higher incidence of IPD in care homes or in immunocompromised people.

Further, the additive structure of the model allows us to quantify the contribution of multiple viruses to the IPD counts, and at the same time the multivariate age-specific model allows a better characterisation of each of these interactions. Finally, the endemic component captures considerable proportions of IPD incidence in all age groups. We can think of this seasonal background as the proportion of disease probably due to some common environmental factors. The adequacy of temperature and rainfall observations to replace trigonometric functions, supported by enhanced model fit both at aggregate and age-specific level, reinforces this hypothesis and allows relaxing the assumption of fixed periodicity, with similar amplitude and timing across seasons. The appropriateness of shared coefficients for rainfall also suggests that disease seasonality has similar timing across the entire population.

As described in section 1.5, the data used might hide some biases. Despite integrating primary care data on ILI with results of virological testing, our model still relies on the

Age	endemic	Influenza	Rhinovirus	RSV	IPD same age	IPD other age
<5	50.35 (34.23-66.91)	0.00	4.49 (<0.01-12.20)	1.31 (<0.01- 5.26)	26.32 (16.24-33.95)	17.53 (3.28-32.61)
5-14	49.94 (21.31-75.08)	8.54 (4.21-13.43)	0.00	0.00	15.68 (5.06-24.10)	25.84 (1.65-55.97)
15-44	26.35 (15.85-38.35)	3.56 (1.69- 5.82)	0.00	0.00	19.40 (10.59-26.88)	50.70 (38.19-63.20)
45-64	29.24 (20.87-39.76)	0.91 (<0.01- 2.83)	5.43 (2.23-8.91)	4.18 (1.58-6.91)	17.15 (8.04-24.27)	43.09 (32.64-53.07)
65+	29.05 (21.65-38.27)	0.00	5.68 (3.03-8.32)	3.91 (1.83-6.38)	23.75 (14.97-30.68)	37.62 (29.41-46.45)

Table 4.10 Model K: Relative proportions (%) of IPD cases attributed to pneumococcal transmission within and across age groups, to influenza, rhinovirus and RSV

assumption that viral surveillance is consistent over time and adequately represents the true burden in the population [105]. Further, we simply multiply the proportion of positive samples by the ILI rates, whereas a joint modelling approach would take uncertainty into account. In terms of IPD data, we believe that testing policies must be consistent over time due to the life-threatening nature of such a condition, and that reporting along UK-wide guidelines [51] was relatively stable over time. Nonetheless, the limited numbers of cases, especially in the age-specific analysis, made the resulting estimates uncertain.

Despite our efforts to mimic disease mechanisms, a number of assumptions are made in our analysis that might be inaccurate or introduce some bias. Linearity between disease incidences is assumed, however interaction between pathogens could be resulting from more complex nonlinear dynamics. The assumption of one week lag between events is the best approximation given weekly data, however the infectious time might be shorter than that [131]. Autoregressive coefficients fixed over time keep our model easy to interpret and avoid overfitting, however such an assumption implies that both pneumococcal transmission and its interaction with influenza have no seasonal behaviour; as a consequence, any season-specific variation is included into the endemic component, summarising unknown aspects such as climatic influence on disease susceptibility. The use of age-structured contact patterns leads to improved model fit compared to an assumption of random mixing between age groups (results not shown), nonetheless the used contact patterns are approximated by a matrix estimated in 2005-06 [149]. Finally, we assume pneumococcal infection to follow from transmission, yet we are aware that pneumococcus can be carried asymptomatically, and that those individuals also contribute to the transmission.

Chapter 5

Time-series methods to assess the impact of an intervention

5.1 Introduction

Public policies or interventions are implemented to improve the health of the public, and the extent to which they achieve their aim should be measured using suitable statistical methods. Quantifying the effects of a policy is of primary importance not only to make a decision on its continuation. Evaluation of an intervention can be also an important ex-ante policy impact assessment tool, as it can provide evidence for similar policies that might be implemented in similar populations in the future.

Assessing the impact of an intervention is often a non-trivial task. An adequate policy evaluation analysis generally involves measuring the resulting outcome of a policy and comparing with the expected outcome in the absence of any intervention. But how should the expected outcome be defined? Let us consider the example of publicly-provided vaccination, offered to healthy individuals in order to prevent infections and mortality: should the outcomes of the participants be compared to their pre-intervention situations, or instead, compared with those of the non-participants [169, 37] ?

Several methodologies have been proposed to identify the intervention effects [122], with the intervention evaluation literature gaining increasing importance in recent years. This chapter introduces a general framework for the evaluation of public health interventions and describes some of the methods currently used to measure their impact.

5.2 Intervention evaluation framework

Our discussion focuses on measuring the causal effect of a non-randomised binary intervention on an outcome of interest, Y. We adopt the potential outcomes framework [179], also known as the Rubin causal model [92]. Under this model, for each treated unit i there are two potential outcomes, $Y_i^{(0)}$ and $Y_i^{(1)}$. $Y_i^{(0)}$ represents the outcome that would be observed if intervention were not applied, and $Y_i^{(1)}$ is the outcome that would be observed if the intervention were applied. The intervention's causal effect is defined as the difference between these two quantities. The fundamental evaluation problem is that we cannot observe simultaneously the same unit in the two scenarios; the scenario in which the intervention is not implemented, the counterfactual, is always missing. Therefore, evaluating the impact of an intervention involves in the first place estimation of the counterfactual [122].

For the scope of this thesis we focus on the situation where the outcome of interest is measured at several time points, before and after the intervention, and the intervention is implemented at population level. Hence, an aggregate entity (e.g. hospital, city, region) is the treatment unit, and accordingly, the outcome is reported in the aggregate form of a time series Y_t, for $t \in \{0, 1, \ldots, T\}$ (e.g. disease incidence). We assume the intervention is introduced at a known time $t_1 \leq T$. The counterfactual $Y_t^{(0)}$ is then estimated for $t > t_1$, and the causal effect of the intervention is defined as $\hat{\tau}_t = Y_t^{(1)} - \hat{Y}_t^{(0)}$. Given the widespread availability of aggregate data and the fact that many policy interventions take place at population level, examples of studies modelling the counterfactual for aggregate outcomes can be found across research areas, from public health (e.g. mortality rates [180]) to economics [101] and sociology (e.g. crime rates [182]). In the next sections we present selected statistical methods that have been developed to model the counterfactual in the described setting.

5.3 Before-and-after designs

Before-and-after designs approximate the counterfactual $Y_t^{(0)}$ by using the outcome observed in the same population in the pre-intervention period, Y_t for $t \leq t_1$, as a control. While such a comparison might sound naïve, it has one main advantage: the studied population acts as its own control, hence time-invariant factors are controlled by design and there is no need to adjust for differences between groups [123].

5.3.1 Interrupted time series analysis

When multiple pre-intervention observations are available, a before-after design is the interrupted time series (ITS) regression model:

$$g(Y_t) = \alpha + f(z_{t_1}) + \kappa\, t + s(t) + \beta \mathbf{X}_t + \varepsilon_t \tag{5.1}$$

where:

- $g(\cdot)$ is an appropriate link function
- α is an intercept term, potentially including the population offset
- $f(z_{t_1})$ defines the impact model, dependent on the binary intervention indicator z_{t_1}, which is function of time t_1
- $\kappa\, t$ and $s(t)$ are the linear trend and seasonal components
- $\beta \mathbf{X}_t$ defines contributions of other time-varying factors X_{jt}

While the intervention is expected to "interrupt" the level and/or trend of the observed outcome $Y_t^{(1)}$ in the post-intervention period, the counterfactual $Y_t^{(0)}$ is estimated by assuming the pre-intervention seasonality, linear trend and relationships of covariates remain unchanged [123, 225]. In summary, this model allows making a pre-post comparison while accounting for observed time-varying factors in the population, i.e. explicitly modelling the contribution of underlying trends and measured time-varying confounders.

While there is general agreement on defining baseline trends as simple means or linear functions of time, specifying the form of $f(z_{t_1})$ can be a less-straightforward choice. Further, the fact that such a choice must be made a priori makes the model inflexible to account for data-driven evidence. Finally, history bias remains a threat, i.e. other events occurring around the same time as the intervention of interest (cointerventions) or unobserved time-varying confounders might affect the outcome, but not be captured by the trend model [123]. Hence, attributing to the intervention any unaccounted change in the outcome might be unreasonable. On the other hand, even when no apparent change in the outcome is observed, the intervention's effect might have been confounded by unobserved factors [118]. To address this limitation, methodologies for counterfactual analysis presented next will focus on controlled designs.

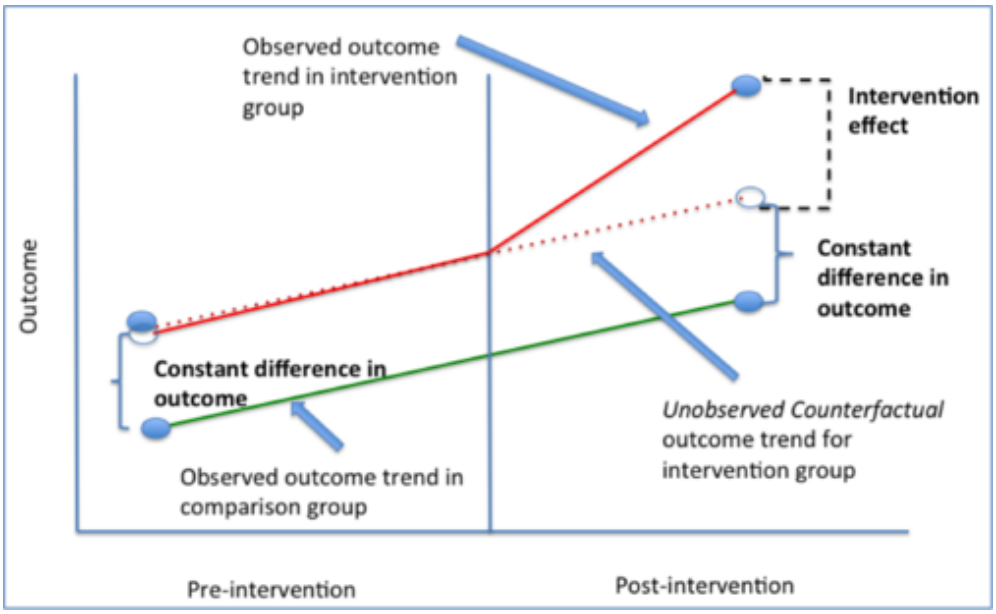


Fig. 5.1 Difference-in-difference estimation, graphical explanation [Columbia University]

5.4 Controlled designs

Methodologies for counterfactual analysis including a control group span across several families of controlled designs, including randomised controlled trials as well as also observational studies [178]. Controlled studies have greater potential to make strong causal statements, as a lack of effect in a well-chosen control can provide stronger evidence to support a causal relationship between the intervention and outcome [123]. For example, the difference-in-difference (DID) method [235] compares the outcome of a treated and a control unit in two time periods (pre- versus post-intervention), assuming that the average outcomes of control and treated units in the absence of an intervention would follow parallel trends (Figure 5.1). The mean change in the control group's outcome from the pre- to the post-intervention period is then used as the counterfactual against which the mean change in the treatment group is evaluated.

Control groups should be chosen to be as similar as possible to the treated group, except for exposure to the intervention [179]. Specifically, controls should not be indirectly affected by the intervention of interest (assumption of no interference), nor exposed to other interventions or events impacting on the control series alone. They can be selected based on a different geographic area, a population subgroup not targeted by the intervention (e.g. gender, age group), or again another outcome unaffected by the intervention. While randomised studies ensure that any differences between groups are at random, randomisation is mostly

infeasible in public health interventions. In observational settings treated and untreated units can be inherently different, particularly when the units can select themselves into participation or when the intervention is targeted towards units at higher risk. Attempts to take differences between groups into account include adjusting for fixed measurable confounders via regression or matching, yet residual confounding due to unknown variables remains unaccounted for [11].

5.4.1 Synthetic controls

In practice, it is often difficult to find a single untreated unit that approximates the most relevant characteristics of the unit exposed to the intervention, and there is often some degree of ambiguity about the chosen measures of affinity between treated and untreated units. As an alternative, multiple control units can be identified. Abadie et al. [1] advocated the use of data-driven procedures to construct suitable control groups, proposing the idea that a combination of untreated units could provide a better comparison for the treated unit than any untreated unit alone. These so-called synthetic control methods combine multiple controls by identifying the weighted average of untreated units that best reproduces characteristics (including observations of the outcome) of the treated unit before the intervention. These weights are then used to estimate the counterfactual in the post-intervention period. Suppose that we observe $J+1$ units, and Y_j is the outcome observed for unit j

- unit 1 is exposed to the intervention at time $t_1 \leq T$
- the remaining J units are untreated units, potential controls
- X_1 is a (k×1) vector of pre-intervention characteristics for the treated unit and $X_{2:(J+1)}$ is a (k × J) matrix which contains the same variables for the untreated units.

If $w = (w_2, \ldots, w_{J+1})$ is a collection of weights such that $\sum_{j=2}^{J+1} w_j = 1$, with each $w_j \geq 0$, then $w^* = (w_2^*, \ldots, w_{J+1}^*)$ is chosen to minimise the distance $||X_1 - X_{2:(J+1)}w||$, commonly evaluated by a weighted Euclidean norm. Then, the synthetic control estimator of the effect of the intervention for the treated unit in a post-intervention period is

$$\hat{\tau_1} = Y_1 - \sum_{j=2}^{J+1} w_j^* Y_j$$

i.e. the difference between the observed and the counterfactual.

5.4.2 Controlled interrupted time series (CITS)

As for before-and-after designs, our focus is on the situation where multiple pre-intervention observations are available. CITS, also called comparative time-series design, are an extension of ITS that model the outcome of the treated unit, $Y_{t,1}$, while considering outcomes of one or more contemporaneous control groups $Y_{t,j}$ for $j = \{1,\ldots,J\}$. Controls should be exposed to any co-interventions or events that affect the treated group: the simultaneity guarantees that any time-specific confounders affecting outcomes of both treated and control units are controlled for. However, controls should not be indirectly affected by the intervention of interest, nor exposed to other interventions or events impacting on the control series alone.

The counterfactual is now defined both on a before-after and on a treated-control comparison. Depending on the situation, different models can be formulated. A single model including indicator variables for the intervention or control series as interaction terms, for example, provides a test of the differential effects of the intervention (level or slope change) across the groups [123]. In the simplest case, a generalisation of the DID seen above, one parametric linear model for the outcome of units $j = \{1,\ldots,J\}$ is formulated assuming that the treatment and control groups follow parallel trends:

$$Y_{j,t} = \mu_t + \kappa_j + \tau_{j,t} d_{j,t} + \beta \mathbf{X_{t,j}} + \varepsilon_{j,t} \tag{5.2}$$

where:

- μ_t describes the common trend and seasonal components
- κ_j is a fixed effect of unit j
- $d_{j,t}$ is the binary intervention indicator, positive for $t > t_1$ in the treated unit $j = 1$
- $\tau_{j,t}$ is the intervention effect for $j = 1$
- $\beta \mathbf{X}_{t,j}$ adjusts for group-specific time-varying covariates $X_{t,j}$

In this case, if a change over time is detected in the intervention group but not in the control, then such an effect is attributed to the intervention. In other words, the intervention effect is evaluated by looking at whether the treatment group deviates from the baseline levels by a greater amount than the controls group(s). More complicated models can be formulated to account for differential changes in covariates between the treated and the control series throughout the study period [186, 123]. However, the more complex the trend, the more difficult it becomes to differentiate intervention effects from natural underlying fluctuations.

5.4.3 The Causal Impact method (CIM)

When multiple controls are considered, synthetic control approaches can be applied to CITS studies, with the general idea of obtaining a weighted average of the controls such that their characteristics are as similar as possible to the ones of the study group. Taking ideas from the synthetic control framework, Brodersen et al. [15] introduced the CIM. They model the outcome of the treated unit, $Y_{t,1}$, including a time-series component that relates $Y_{t,1}$ to previous outcomes on the same unit, $Y_{t-p,1}$, and a regression component that uses the contemporaneous outcomes on control units as covariates. Furthermore, instead of performing a constrained maximisation to match covariates as in section 5.4.1, Brodersen et al. [15] make use of a 'Bayesian structural time series' (BSTS) model including a Bayesian variable selection technique to choose and weight the control outcomes $Y_{t,2:(J+1)}$. In the next sections, we briefly introduce the BSTS model and variable selection in general, before specifying the CIM.

Structural time series models

Structural time series models, also called dynamic linear models (DLM), are univariate time series models widely used in the econometric literature. They describe the evolution of a time series as the sum of a number of independent components, typically a trend component and a random component; a seasonal component is often added. Compared to the static regression models presented in chapter 3, where parameters were fixed for all time, here each component is allowed to evolve randomly over time: a model is specified to decompose it into the sum of an unobservable state variable and a random error. The general DLM formulation, for $t = 1, 2, \ldots, T$, consists of a set of two equations, called the observation and evolution equation respectively:

$$\begin{aligned} Y_t &= F_t \theta_t + \varepsilon_t \\ \theta_t &= G_t \theta_{t-1} + \eta_t \end{aligned} \tag{5.3}$$

where Y_t is the observed outcome at time t; θ_t is an unobservable (latent) state vector representing the outcome trend; and F_t are possibly time-dependent coefficients for the latent state. Such a latent state is itself dynamically modelled through the evolution equation: the evolution matrix G_t deterministically maps the parameter space from one time step to the next, so the states at time t are temporally related to those before and after. A seasonality component γ_t can also be added to the observation equation, and a corresponding evolution equation specified for it.

Both equations include error terms, ε_t being the observation error and η_t the evolution error, which are assumed mutually independent. DLMs are a special case of state-space models, a general class of non-stationary time series models [168], which assume linearity and Gaussianity of errors. Generalised DLMs relax the assumption of normality by allowing the distribution to be any of the exponential family of functions (typically Bernoulli, binomial and Poisson distributions, useful in particular for count data).

Bayesian variable selection via the spike-and-slab prior

When estimating time series models over relatively short time periods with multiple variables and large numbers of parameters, multicollinearity and overfitting can be problematic. As an alternative to regularisation techniques such as Lasso and Ridge regression [60], a Bayesian approach allows imposing priors on the regression coefficients to select the predictors in a robust and automatic way. Here we focus on the **spike-and-slab prior** approach to sparsity in Bayesian variable selection [145, 65, 66].

Let δ denote a vector, of the same length as β, that indicates whether or not a particular covariate is included in the regression. In formulae, $\delta_i = 1$ indicates $\beta_i \neq 0$ and $\delta_i = 0$ indicates $\beta_i = 0$. Let β_δ indicate the subset of β for which $\delta_i = 1$, and let σ^2 be the residual variance from the regression model. A spike-and-slab prior for the joint distribution of $(\beta, \delta, \sigma^2)$ can be factorised in the usual way:

$$p(\beta, \delta, \sigma^2) = p(\beta_\delta \mid \delta, \sigma^{-2}) p(\sigma^{-2} | \delta) p(\delta).$$

- for the "spike" part of a spike-and-slab prior, which refers to the point mass at zero, a Bernoulli distribution for each i is generally assumed, so that the prior is a product of Bernoullis:

  $$\delta \sim \prod_{i=1}^{K} \pi_i^{\delta_i} (1 - \pi_i)^{1-\delta_i}$$

  When detailed prior information is unavailable, it is convenient to set all π_i equal to the same probability, π. The common prior inclusion probability can easily be elicited from the expected number of nonzero coefficients. If k out of K coefficients are expected to be nonzero, then set $\pi = k/K$ in the prior.

- The "slab" component is a prior for the values of the nonzero coefficients, conditional on knowledge of which coefficients are nonzero. Let b be a vector of prior guesses for regression coefficients, let Ω^{-1} be a prior precision matrix, and let Ω_δ^{-1} denote rows

and columns of Ω^{-1} for which $\delta_i = 1$. A conditionally conjugate "slab" prior is

$$\beta_\delta \mid \delta, \sigma^2 \sim N\left(b_\delta, \sigma^2(\Omega_\delta^{-1})^{-1}\right)$$

$$\frac{1}{\sigma^2} \sim \Gamma\left(\frac{df}{2}, \frac{ss}{2}\right)$$

It is conventional to assume $b = 0$ (with the possible exception of the intercept term) and $\Omega^{-1} \propto X^T X$. The final values that need to be chosen are df and ss. These can be elicited as a function of the R^2 statistic you expect to obtain from the regression, and the weight you would like to assign to that guess, measured in terms of the equivalent number of observations. The df parameter is the equivalent number of observations, and $ss = df(1-R^2)\sigma_y^2$.

This leads to a posterior distribution with positive mass at zero for sets of regression coefficients, while Bayesian model averaging smooths the predictions over a large number of potential models, as well as providing the posterior inclusion probability for each predictor.

The CIM as a BSTS

The BSTS model proposed for the CIM by Brodersen et al. [15] takes the form

$$\begin{aligned} Y_{t,1} &= \mu_t + \gamma_t + X_t\beta + \varepsilon_t \\ \mu_t &= G_t\mu_{t-1} + \eta_t \\ \gamma_t &= \sum_{j=1}^{s/2} \gamma_{j,t} \end{aligned} \tag{5.4}$$

where

- $Y_{t,1}$ is the observed outcome in the treated unit $j = 1$;
- μ_t is the latent state, i.e. $F_t = 1$ and $\theta_t = \mu_t$ in equation 5.3;
- γ_t is a seasonal component, which can be expressed via trigonometric functions according to the periodicity s, using two time-varying harmonic components $\gamma_{j,t-1}$ and $\gamma^*_{j,t-1}$ evolving through time as

$$\gamma_{j,t} = \gamma_{j,t-1}\, cos\left(\frac{2\pi j}{s}\right) - \gamma^*_{j,t-1}\, sin\left(\frac{2\pi j}{s}\right) + \omega_{j,t}$$

$$\gamma^*_{j,t} = \gamma^*_{j,t-1}\, cos\Big(\frac{2\pi j}{s}\Big) - \gamma_{j,t-1}\, sin\Big(\frac{2\pi j}{s}\Big) + \omega^*_{j,t}$$

- $X_t\beta$ is a static regression component with time-varying covariates X_t and time-fixed coefficients β. Specifically, we define $X_t = Y_{t,2:(J+1)}$, i.e. the $X_{2:(J+1)}$ matrix of section 5.4.1 is now time-dependent and contains J columns consisting of the observed outcomes $Y_{t,2:(J+1)}$ in potential control units. These candidate predictors are selected and combined into a single synthetic control based on their ability to predict the outcome of the treated unit prior to the intervention, via a spike-and-slab prior for β;
- ε_t, η_t and ω_t are incorrelated Gaussian error terms.

Several choices can be considered for the evolution matrix G_t, which describes the evolution of the latent state vector μ_t from one time step to the next. The most commonly used state models are:

- local level model, i.e. a random walk

$$\mu_t = \mu_{t-1} + \eta_t;$$

- local linear trend, i.e.

$$\mu_t = \mu_{t-1} + \phi_t + \eta_t,$$
$$\phi_t = \phi_{t-1} + \upsilon_t,$$

  with an additional error term υ_t;

- autoregressive model of order p, AR(p), specifying that μ_t depends linearly on its own previous p values

$$\mu_t = \sum_{i=0}^{p-1} \rho_i \mu_{t-i} + \eta_t$$

  – for large p, a spike and slab prior could be applied on the autoregression coefficients, so that some coefficients might be set to zero.

After fitting the model to the observed data $y_{t,1}$ for $t \leq t_1$, the counterfactual $y^{(0)}_{t,1}$ is forecast for the post-intervention period $t > t_1$ assuming that the relationship between the treatment and the control series (equation 5.4) that existed prior to the intervention continues afterwards. Thus, the observed outcomes for the control units $y^{(0)}_{t,2:(J+1)}$ in the post-intervention period are employed to predict the counterfactual $y^{(0)}_{t,1}$, and a posterior distribution for the causal

effect at each time point $t > t_1$ is then obtained by the posterior samples of $y_{t,1} - y_{t,1}^{(0)}$.

5.5 Predictive model assessment

As anticipated in section 3.3, model assessment is essential to quantify the goodness of fit but also to identify the best performing model among a set of choices. Ideally we should check the performance of our model on new data, as evaluating the model on the data used to train it would lead to a biased estimate of performance. In practice, new data are often not available, and a new dataset can be artificially created by splitting the available data in two portions: if we call y^{fit} the portion of y used to fit the model, y^{crit} can denote the portion used for model criticism.

Cross-validation is a standard way to split the dataset: when independence between observations holds, the $k\%$ of the data left out from y^{fit} is chosen at random. In time series settings, due to the inherent temporal dependence, splitting of the data must respect the temporal order in which values are observed: the training set can consist of observations up to a given time point y_k, and any future observation with respect to y_k can form the test set. Depending on the situation, forecasting performance can be assessed by looking at one-step forecasts or multi-step forecasts. This approach, often referred to as **out-of-sample prediction**, provides a good proxy for how the model will perform in a real world forecasting environment, in which we stand in the present and forecast the future [9].

To obtain a more robust estimate of the expected performance of the chosen method, the process of splitting the time series into train and test sets can be repeated by choosing different splitting points, varying the number of records used to train the model. Typical splits are 50-50, 70-30 and 90-10, however the minimum number of observations required to train the model must be carefully chosen so that y^{fit} is large enough to be representative of the original problem. Moreover, it is worth noting that different lengths of the test set must be taken into account in order to obtain performance statistics that can be meaningfully combined and compared. For example, forecast accuracy can be computed by averaging over the test sets. This procedure is sometimes known as “evaluation on a rolling forecasting origin” because the “origin” at which the forecast is based rolls forward in time. Some authors also talk about nested cross-validation, as this comes with an additional computational expense of training and evaluating multiple models, selecting parameters that minimise the prediction error for

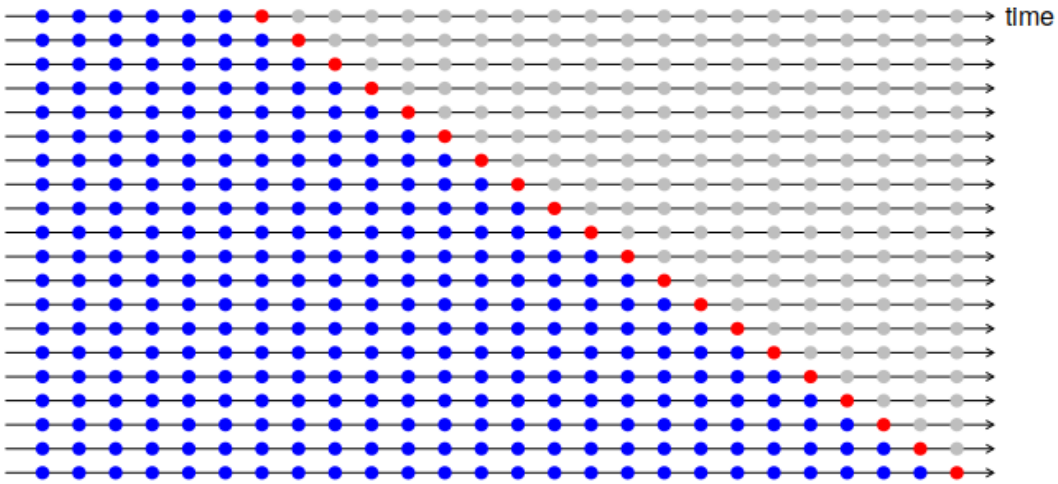


Fig. 5.2 Graphical representation of evaluation on a rolling forecasting origin

.

each choice of training-test set [9].

In a Bayesian setting, model criticism typically analyzes the **posterior predictive distribution** [63]: instead of simply considering a point estimate for $\hat{\mathbf{y}}^{\text{crit}}$, typically its mean, we look at the distribution of unobserved data $\mathbf{y}^{\text{crit}}$, conditional on the observed values $\mathbf{y}^{\text{fit}}$. It is an average of conditional predictions of the new data, averaged over the posterior distribution of parameters θ, so that uncertainty about θ is taken into account:

$$p(\mathbf{y}^{\text{crit}} \mid \mathbf{y}^{\text{fit}}) = \int p(\mathbf{y}^{\text{crit}} \mid \theta, \mathbf{y}^{\text{fit}}) p(\theta \mid \mathbf{y}^{\text{fit}}) \mathrm{d}\theta.$$

Comparison between observed and predicted values can be done using multiple checking functions. We consider here the Bayesian Mean Squared Prediction Error (MSPE)

$$\int (\mathbf{y}^{crit}_{pred} - \mathbf{y}^{crit})^2 p(\theta \mid \mathbf{y}^{\text{fit}}) \mathrm{d}\theta$$

where $\mathbf{y}^{crit}_{pred}$ is a vector of replicates drawn from $p(\mathbf{y}^{\text{crit}} \mid \mathbf{y}^{\text{fit}})$.

5.6 Conclusions

In this chapter we have presented two classes of methods that have been used to model intervention effects, playing a prominent role in identifying the causal effects of interventions

in many institutions and international organisations. Each of them has advantages and limitations, in particular the ITS relies on a pre-chosen function for the intervention effect, whereas the CIM is more flexible but nevertheless requires specification of priors for the models.

We explore their application in the next chapter, investigating the impact of pneumococcal vaccine introduction in England. This adds to existing evidence produced by PHE, that uses the counterfactual analysis somewhat parsimoniously in its evaluation and ex-ante policy impact assessment guidelines, relying on simple impact indicators [113].

Chapter 6

Application to Pneumococcal Conjugate Vaccine

6.1 Introduction

Streptococcus pneumoniae is an important cause of severe infection and death worldwide, especially in children: according to the Word Health Organisation it is associated with about 5% of the global infant mortality [234]. Since resistance of *S.Pneumoniae* to multiple classes of antibiotics has increased in recent years [21, 221], prevention of disease via pneumococcal immunisation has become a public health priority [29]. Two classes of pneumococcal vaccines are available: pneumococcal polysaccharide vaccines (PPVs) and pneumococcal conjugate vaccines (PCVs). The first PPV version was licensed as early as 1946 in the United States [127], however the type of antibody response that it induces (T-cell-independent) [6] makes it ineffective in children younger than 2 years of age. Hence, attention has shifted towards PCVs when planning infant immunisation programs, since PCVs can provide protection against nasopharyngeal carriage and disease in vaccinated individuals of any age, and consequently reduce the overall transmission of the pneumococcus, leading to herd immunity [111].

S.Pneumoniae, as many other disease-causing bacteria, coats itself with a polysaccharide capsule to hide from the human immune system. Since the early 20^{th} century it has been shown that the type of capsule determines the virulence and propensity to cause invasive disease [47], to such an extent that a famous immunology book states: "from the point of view of the adaptive immune system, each serotype of *S.Pneumoniae* represents a distinct organism" [106]. Over 90 capsule serotypes have been identified as of today [90, 64], and

efforts have been put into developing vaccines that could prevent disease caused by the most clinically relevant serotypes. A seven-valent formulation first became available in 2000 and included the seven serotypes most commonly isolated in children under 5 years in the pre-PCV era, while PCV10 and PCV13 were developed soon after [58].

Efficacy of these vaccines has first been assessed in pre-licensing randomized clinical trials: up to 97% IPD reduction was observed in the first vaccinated individuals [10, 158]. However, an increase in carriage due to some non-vaccine serotypes (NVT) has also been observed since the early PCV trial phase [157, 130, 41], reducing the benefits of vaccination. This happened because protection against nasopharyngeal carriage for serotypes included in the vaccine, an advantage of PCVs compared to PPVs, opened an ecological niche that altered carriage epidemiology, a phenomenon known as *serotype replacement*.

Since PCVs have become part of the routine infant immunization schedule in several high income countries [59], further work in ecological settings has confirmed the decline in pneumococcal disease burden in terms of bacteraemia, pneumonia and otitis media [10, 171, 7, 188, 163]. Nonetheless, a precise characterisation of serotype replacement would be helpful to clarify past changes in pneumococcal disease epidemiology and to inform the likely impact of future vaccines containing additional serotypes (higher valency). A few studies have investigated the magnitude of serotype replacement in carriage, showing that the total prevalence of pneumococcal carriage was left unchanged due to complete replacement by pneumococci expressing NVT [57, 42]. Likewise, most surveillance studies assessing the impact of PCVs on serotype-specific IPD identified a reduction in the number of infections caused by vaccine serotypes (with the exception of serotype 3 [119, 113]) and increased rates of NVT disease due to the high invasiveness potential of some NVTs of increased carriage [57].

Yet, the estimated magnitude of serotype replacement varied from country to country: in England, Ladhani et al [113] recently concluded that *"6 years after the introduction of PCV13, the additional benefits of this higher-valent vaccine have been nearly abolished by replacement disease"*, while the USA [117] did not identify the substantial NVT increase observed in other countries [215, 223, 7, 77]. Changes in surveillance practices, transmission dynamics, population risk factors, and pathogen evolution have all been speculated to play a role in such differential replacement disease across populations [216], nonetheless none of these factors individually appears sufficient to account for the observed differences.

We believe that the significant challenge of choosing sound statistical methodology to evaluate the impact of PCVs introduction may also be playing an important role in finding the disagreement over quantification of serotype replacement. Simplistic before-after models have often been employed in previous work, producing incidence rate ratios (IRRs) to summarise changes in yearly incidence of serotype-specific IPD after PCV introduction [142, 226, 113]. Firstly, despite availability of multiple pre-intervention observations, occurrences of incidence were assumed to be independent and no time-series component was used, hence underlying trends were not adequately modelled. Secondly, aggregating incidence to year level might result in irregular patterns of difficult attribution, whereas modelling monthly counts allows a finer detail of time-varying covariates. Finally, no control group was considered: these type of analyses assume that no factors other than the intervention might have affected the outcome of interest, i.e. they attribute any change in the outcome to the vaccine introduction, completely neglecting the importance of time-varying confounding.

An evaluation method that makes use of control time series to model counterfactuals would produce a more reliable inference: Thorrington et al. [205] first estimated changes in incidence of pneumonia, sepsis and otitis pre-PCV and post-PCV comparing them to changes in incidence of a composite control, obtained by calculating the geometric mean of the IRRs of five control conditions. Yet, an integrated time series approach where selected controls for IPD are combined using synthetic controls methodology instead of a geometric mean would be more robust to the choice of controls.

Following the work of the previous chapters, where we estimated the impact of seasonal and pandemic flu on IPD at a time when PCV7 and PCV13 were introduced in England and Wales (2006 and 2010 respectively), we use serotype-specific IPD incidence rates to better understand the role of PCV7 and PCV13 in shaping overall IPD trends and serotype replacement in England. We select the period 2000 to 2018 in order to distinctly assess changes related to PCV7 and PCV13 introduction, disentangling their different contributions, and we employ and compare both an ITS model and a synthetic control method to adjust for testing and reporting trends.

6.2 Data

As described in section 4.2, counts of positive isolates for a number of clinically significant pathogens are reported weekly to PHE by all the diagnostic microbiology laboratories included in the national surveillance system, and stored in the SGSS database. Notification

of all confirmed IPD cases has been enforced with the 2010 Health Protection Legislation, whereas reporting was voluntary before 2010. Information is presented in terms of incidence rate per million residents, both at population level and by age group, where we consider five age groups defined as 0-4, 5-14, 15-44, 45-64 and 65+ years old. Information about influenza positivity, temperature and rainfall, as described in 4.2, is also added.

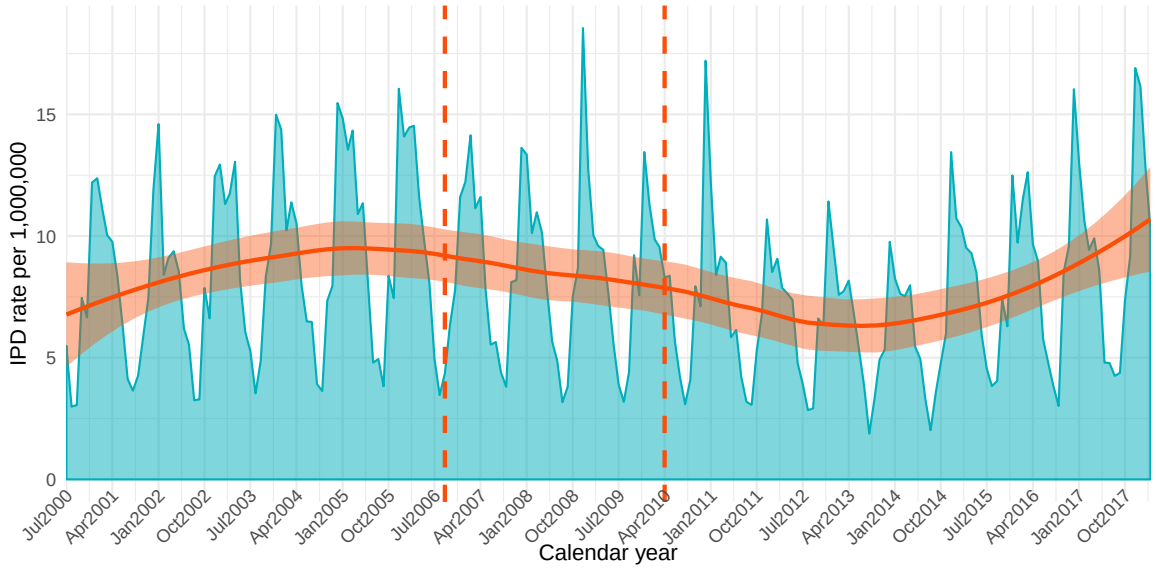


Fig. 6.1 Monthly IPD incidence rate per million residents

.

A total of 96,852 IPD cases have been notified to PHE during the 18 years study period, from July 2000 to June 2018, with a mean incidence rate of 8.1 cases per million residents each month. Figure 6.1 displays the overall temporal trend of monthly incidence rate, characterised by sharp winter peaks not too dissimilar across seasons both in terms of width and timing. However, the amplitude of the peaks saw some variations in the observed period, as summarised by the loess (locally weighted smoothing) function in red: it gradually increased from 2000 to 2006, then decreased until 2013, and surged again in the last five years, overtaking the maximum levels of 2006.

Age-specific IPD incidence is inspected in Figure 6.2, showing very diverse trends across groups. Please note that scale differs across panels. In young children, the vaccinated group, the detected IPD incidence reached a maximum of 219 cases/year per million residents in 2006 and steadily decreased until 2012, when it stabilised at 76 cases/year per million residents until 2018: overall IPD incidence was down to a third within six years from the introduction of PCV vaccination. In school-aged children (5-14) detected IPD incidence

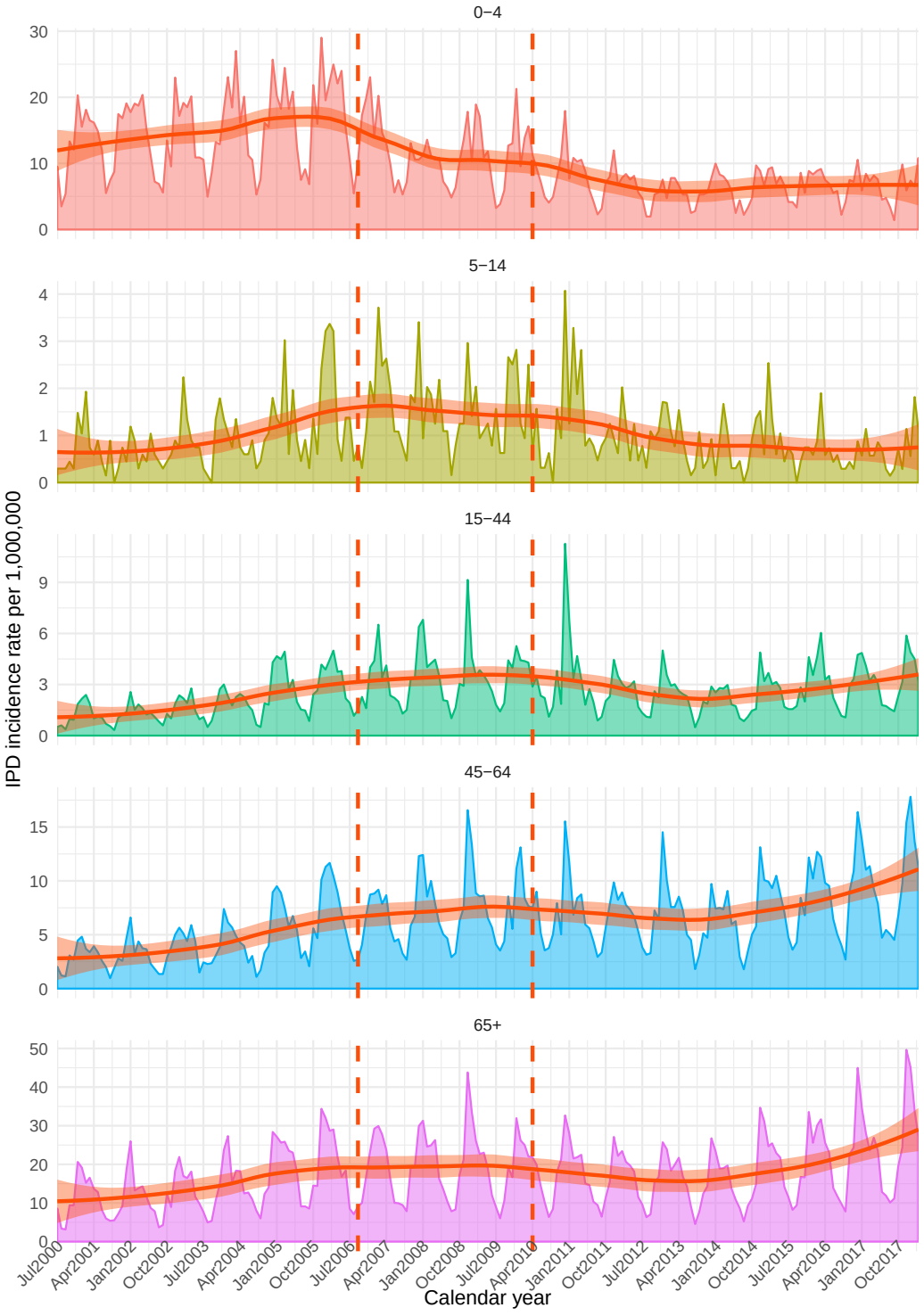


Fig. 6.2 Monthly age-specific IPD incidence rate per million residents

.

is much lower: it peaked at 19.7 cases/year per million residents in 2007 and stabilised at 8.4 cases/year per million residents from 2013 onwards; compared to infants, the decrease kicked in one year later, and the pre-vaccination incidence was halved. In adults aged 15-44 the detected IPD incidence did not decreased until 2011, when it reached a maximum of 45 cases/year per million residents, and after a minimum in 2014-2015 it raised again in 2017-2018 to the same levels of 2007-2008. A similar trend was observed in adults aged 45-64 and 65+: after a modest decrease in 2014-2015, detected IPD incidence showed an upward trend, reaching its highest levels in 2016, 2017 and 2018.

Additionally, serotyping of isolates positive for *S.Pneumoniae* is performed at the Pneumococcal Reference Laboratory at PHE, Colindale, London. We first look at monthly numbers of samples positive for each serotype, and we then aggregate such counts by "PCV group". Incidence for each serotype, divided in three panels according to the PCV group, is presented in Figure 6.3.

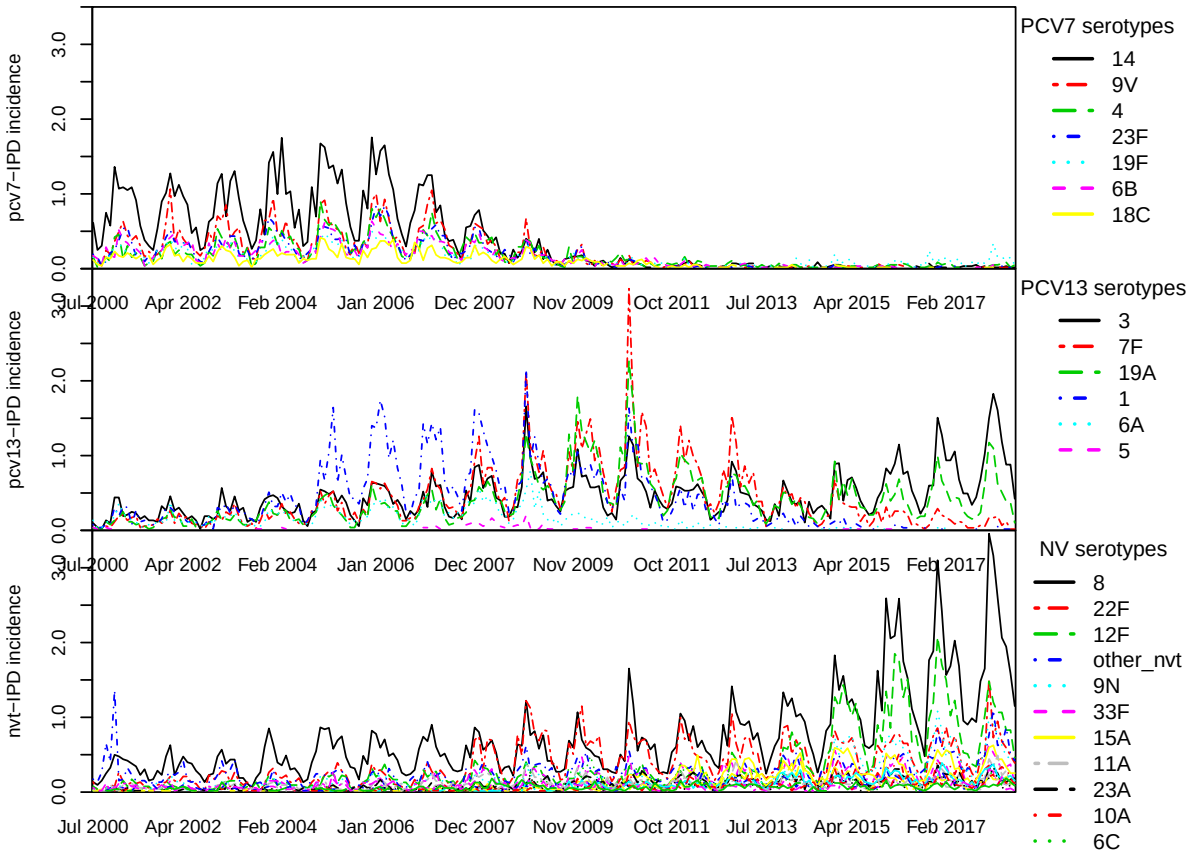


Fig. 6.3 Monthly IPD incidence rate per million residents by serotype

PCV7 serotypes refers to serotypes 4, 6B, 9V, 14, 18C, 19F, and 23F, covered by PCV7 vaccination: these serotypes accounted for 51.3% of IPD cases up to 2006, but they saw a sharp decrease after that, becoming responsible for a negligible number of IPD cases (2.65% of cases in the last five years) within less than two years from vaccine roll-out (top panel of Figure 6.3). *PCV13 serotypes* refers to the 6 serotypes added to PCV13 and not included in PCV7, namely 1, 3, 5, 6A, 7F, and 19A. Incidence of IPD due to PCV13 serotypes (middle panel of Figure 6.3) saw an increase reaching its maximum levels around 2010 and 2011, when it represented 48% of total incidence, and then steadily decreased afterwards (with the exception of serotype 3, for which PCV13 effectiveness is debated [134, 69]). All the remaining serotypes are classified as non-PCV7/13 serotypes (NVTs). In the bottom panel of Figure 6.3 we can observe important increase in the IPD incidence due to several NVT, which in the last five years have been responsible for 79% of total incidence, with serotype 8 alone reaching the highest serotype-specific incidence for the entire study period.

Age-specific incidence by PCV group is presented in Appendix B. In all age groups pcv7-IPD incidence reduced importantly after the introduction of PCV7 (Figure B.1), in a more timely manner in the vaccinated group and with some years of lag in the other groups. Pcv13-IPD (Figure B.2) also saw a modest reduction after PCV13 introduction, with constantly decreasing trends across age groups. Finally, NVT-IPD (Figure B.3) stably increased in all groups from 2000 until 2014, however it seems to have stabilised since then.

6.2.1 Selection of controls

Control time series that received no treatment are critical for obtaining accurate counterfactual predictions of what IPD incidence would have been observed had PCV not been introduced in England, since they allow adjusting for effects of confounders otherwise unaccounted for [15]. In absence of serotype replacement, time series of IPD incidence due to serotypes not affected by the intervention would provide a natural counterfactual. However, in practice, serotype replacement showed that such incidence was indirectly affected by the intervention. Similarly, incidence of disease in unvaccinated age groups is indirectly affected due to herd immunity, so they do not make good counterfactuals either.

We thus consider pathogens other than *S.Pneumoniae*, subject to the same SGSS reporting rules and likely to be identified in similar clinical conditions and testing indications. They are selected a priori with the criteria of not interacting with the pneumococcus and not having been the focus of other public health interventions. After discussing with experts, we

extract from SGSS isolates positive for *Haemophilus Influenzae*, *Klebsiella Pneumoniae*, *Pseudomonas Aeruginosa*, *Staphylococcus Aureus* and *Staphylococcus Coagulase Negative* and *Escherichia Coli*, when identified from normally sterile sites (invasive cases).

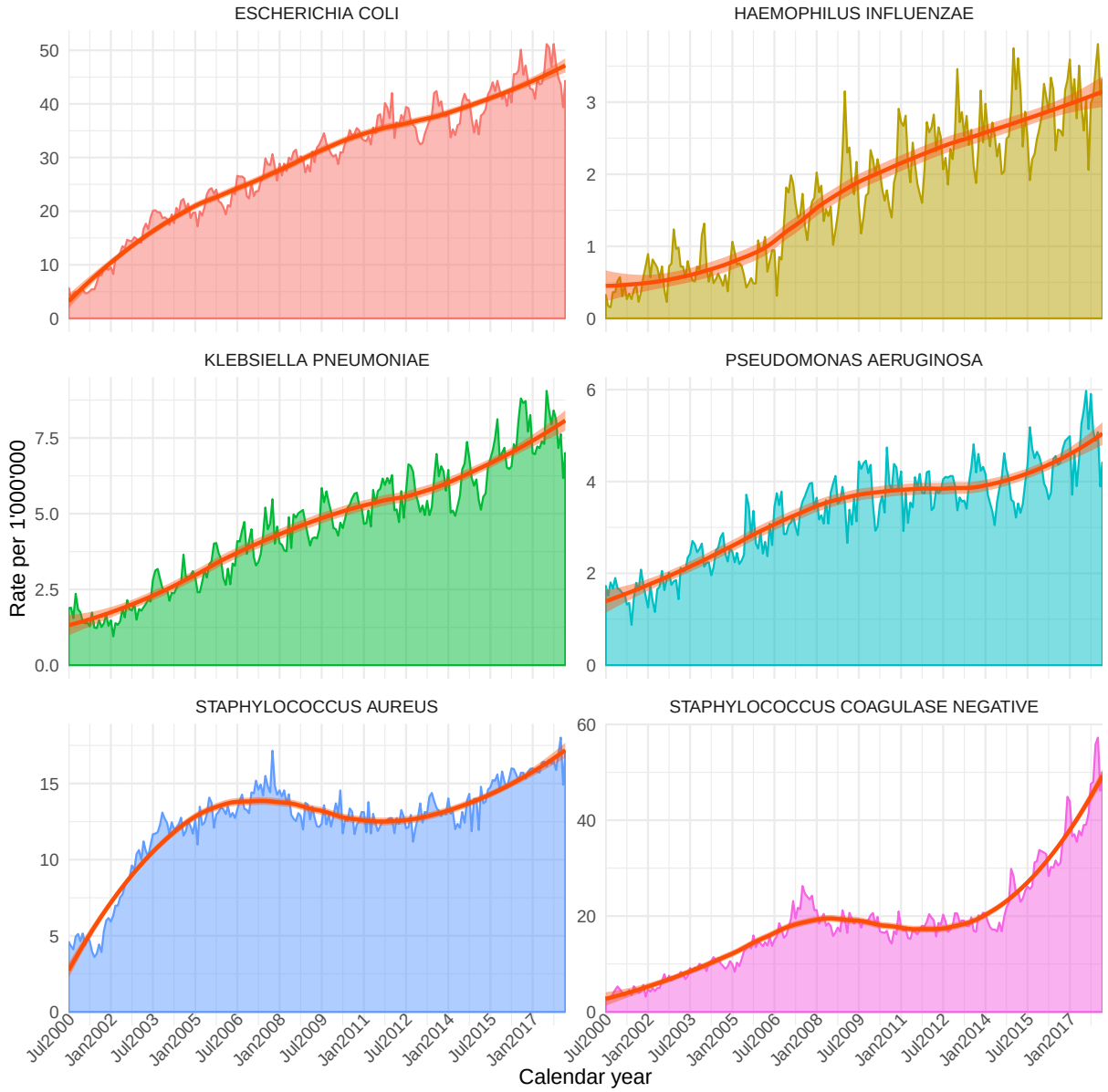


Fig. 6.4 Control time series - incidence rate per million residents

.

Raw trends of recorded invasive bacterial infections in the years immediately before (2000–06) and after (2010-18) introduction of PCV7 and PCV13 are presented in Figure 6.4, using aggregate rates per million residents in England. While IPD rates showed a rather

flat trend, growing modestly until 2006, declining until 2014 and raising again afterwards, national monthly rates of other invasive pathogens increased steadily during this period, registering a 3-fold to 10-fold increase in incidence. Although all trends suggest an increased testing and reporting over time, such increase does not appear to follow a common pattern for the different pathogens. Age-specific incidence for the control time series is plotted in Appendix B, Figure B.4.

6.3 Analysis strategy

6.3.1 ITS regression

We start by considering ITS regression models, with the aim of estimating the impact of PCV7 and PCV13 introduction in September 2006 and April 2010 respectively. Since we are dealing with counts, we use a Poisson regression model: $y_t \sim Poi(\mu_t)$ where

$$\log(\mu_t) = \log(pop_t) + f(PCV_t) + s(t) \tag{6.1}$$

following the general formulation in equation 5.2. In order to characterise the natural time-varying behaviour of IPD incidence, we model the $s(t)$ component as a linear function of influenza, temperature and rainfall observations, that have previously shown to effectively mimic IPD seasonal behaviour in a rigorous model selection in section 3.3. In formulae, $s(t) = \gamma_1\ \text{flu}_t + \gamma_2\ \text{temp}_t + \gamma_3\ \text{rain}_t$. No additional covariates are included.

We explore different model choices for the impact model $f(PCV_t)$: despite not knowing the shape of intervention effect, a limitation of this regression framework is the requirement to specify the impact model a priori, as explained in section 5.3. The simplest model considered is the so-called interrupted intercept model, which assumes that intervention alters the level of incidence by a constant κ_1, but does not affect its trend. In formulae:

$$\log(\mu_t) = \log(pop_t) + \kappa_1 postPCV_t + s(t) \tag{6.2}$$

We then compare it to a model that introduces a time trend and also tests whether such trend differs before and after vaccine introduction, i.e.

$$\log(\mu_t) = \log(pop_t) + \kappa_1 postPCV_t + \kappa_2 t + \kappa_3 t postPCV_t + s(t) \tag{6.3}$$

In order to inspect how results differ across model choices, we compare goodness of fit in terms of likelihood ratio tests since we are looking at nested models: we first test the significance of a time component, and then the relevance of the time-intervention interaction term. In all the models, the counterfactual is obtained by assuming that underlying trends estimated in the pre-intervention period remain the same afterwards, i.e. $\kappa_1 = \kappa_3 = 0$.

The different models considered are summarised in Table 6.1. We first perform this analysis on the IPD time series pertaining the entire population, without serotype stratification, to investigate any change in IPD incidence overall (model A). In the following step, we disentangle the vaccine's direct effect from serotype replacement by slicing IPD incidence in two groups: effect of the vaccine can be quantified by modelling the time series of IPD incidence in the "treated" group, i.e. IPD cases due to serotypes targeted by the vaccine (model B for PCV7, model D for PCV13). Impact of serotype replacement, on the other hand, can be quantified by modelling time series of IPD incidence due to the remaining serotypes (model C for PCV7, model E for PCV13).

name	outcome var	"untreated" period	"treated" period
model A	IPD	Jul 2000 - Sept 2007	Oct 2007 - Jun 2018
model B	PCV7-IPD	Jul 2000 - Sept 2007	Oct 2007 - Apr 2011
model C	nonPCV7-IPD	Jul 2000 - Sept 2007	Oct 2007 - Apr 2011
model D	PCV13-IPD	Sept 2007 - Apr 2011	May 2011 - Jun 2018
model E	NVT-IPD	Sept 2007 - Apr 2011	May 2011 - Jun 2018

Table 6.1 Summary of pre- and post-intervention periods considered for different outcome variables. One year lag was considered between the policy enactment date and the start of the "treated" period, hence we have Sept 2007 instead of Sept 2006 and Apr 2011 instead of Apr 2010.

6.3.2 CIM

The second approach proposed for our analysis involves BSTS models, described in equation 5.4. In their univariate form, they are defined by equations

$$\text{IPD}_t = \mu_t + \gamma_t + \mathbf{X}_t \mathbf{X}_t \mathbf{x}_t \beta + \varepsilon_t \quad \mu_t = G_t \mu_{t-1} + \eta_t \quad \gamma_t = \sum_{j=1}^{s/2} \gamma_{j,t} + \omega_t$$

We model IPD incidence rates by making an assumption of Gaussian distributions for ε_t, η_t and ω_t. The matrix

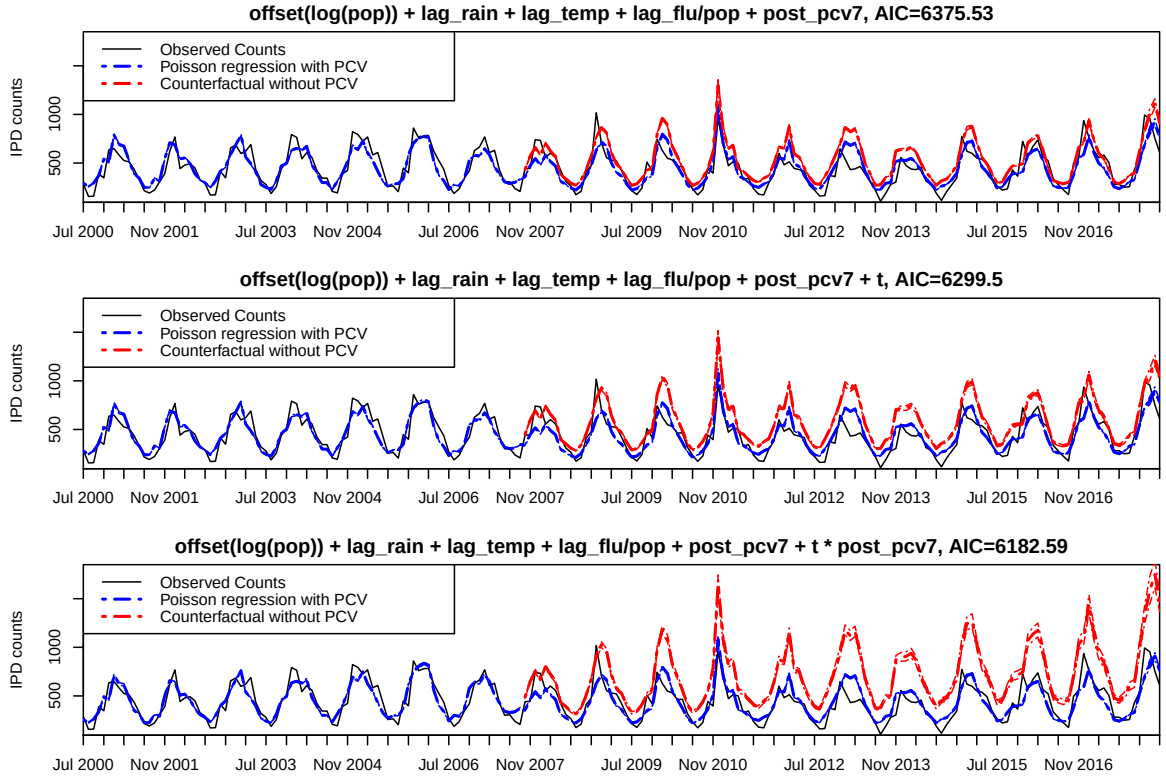


Fig. 6.5 Model A: fitted IPD counts based on three ITS models

6.3.3 ITS analysis

Throughout the ITS analysis, results are plotted using blue lines to display fitted counts under the different impact models and red lines to indicate the counterfactual. In the main text the post-intervention period is considered to start with one year of lag from the intervention date, to allow for one full cohort of children to receive vaccination; a sensitivity analysis on results including a two-year lag is shown in appendix B, Figures B.5, B.6 and B.7.

Results from the regression model on overall IPD incidence are shown in Figure 6.5 (model A): we can see how, under all the three impact models, the observed IPD incidence is lower than the counterfactual, meaning that vaccine introduction had a positive impact on IPD incidence. The percentage of averted cases, i.e. the proportion of cases not observed with respect to the counterfactual, is estimated to be 17.3% (CI 16.1%-18.4%) under the first model, 25.2% (CI 23.2%-27.1%) and 38.8% (CI 33.4%-43.7%), for a total number of averted IPD cases equal to 11702, 18950 and 36138 respectively. As expected, this gap is estimated to be larger when we introduce a time-intervention interaction that allows for an increasing trend until September 2007 and a decreasing trend afterwards. Likelihood ratio

tests suggest that this third impact model also fits the data best.

Further, in order to quantify to which extent the small decrease in IPD incidence is confounded by serotype replacement, we separately model IPD incidence due to PCV7 and non-PCV7 serotypes. For this analysis we need to restrict our observation period up to April 2011, i.e. when we expect the impact of PCV13 to become visible on some of the non-PCV7 serotypes. The top three panels of Figure 6.6 picture the impact on PCV7 serotypes (model B) by model formulation. A large gap between fitted and counterfactual levels for all the three models is estimated, suggesting that they all agree on an important benefit attributable to PCV7 vaccination: estimated proportions of averted cases are 56.6% (CI 54.6%, 58.5%), 65.6% (CI 63.4%-67.6%) and 97.0% (CI 94.3%-98.4%) respectively, with the third model fitting the data best.

On the other hand, when looking at IPD incidence due to non-PCV7 serotypes (model C), different models show disagreement, as represented in the bottom three panels of Figure 6.6: the model in the first panel, that does not include a linear time trend, determines that levels of non-PCV7 IPD have almost doubled, +99.3% (CI 94.4%-104.4%), compared to what would have happened without vaccine. However, when IPD incidence is assumed to linearly increase in time before the intervention, then observed and counterfactual estimates almost coincide: the observed incidence after PCV7 introduction only exceeds by 2.5% (CI -2.0%; 7.3%) what would have happened regardless of vaccine introduction. Once again, the third model fits the data best, estimating a 42.7% (CI 23.8%-57.0%) reduction in non-PCV7 IPD incidence, yet we might question whether we aren't overfitting.

Similarly, the impact of PCV13 introduction is assessed by looking at IPD incidence due to the six additional serotypes covered by PCV13 but not PCV7 (model D): the pre-intervention period is defined from September 2007 (one year after PCV7 introduction), and the counterfactual is modelled from April 2011 onwards, allowing one year lag from PCV13 introduction. Serotype replacement is quantified modelling counts of IPD due to NV serotypes in the same period (model E). Results on the effectiveness of PCV13, reported in the top three panels of Figure 6.7, agree on a significant reduction of PCV13-IPD incidence (model D), with the proportion of averted cases ranging between 47.6% (CI 45.7%-49.4%), 22.2% (CI 17.0%-27.1%) and 51.2% (CI 22.1%-69.5%) respectively. As for the PCV7 analysis, estimation of serotype replacement is the most challenging (model E): the first model estimates a 72.4% increase (CI 67.6%,77.3%) in NVT-IPD incidence, the second model a 8.1% reduction (3.8%,12.3%) and the third one a 35% increase with large uncertainty

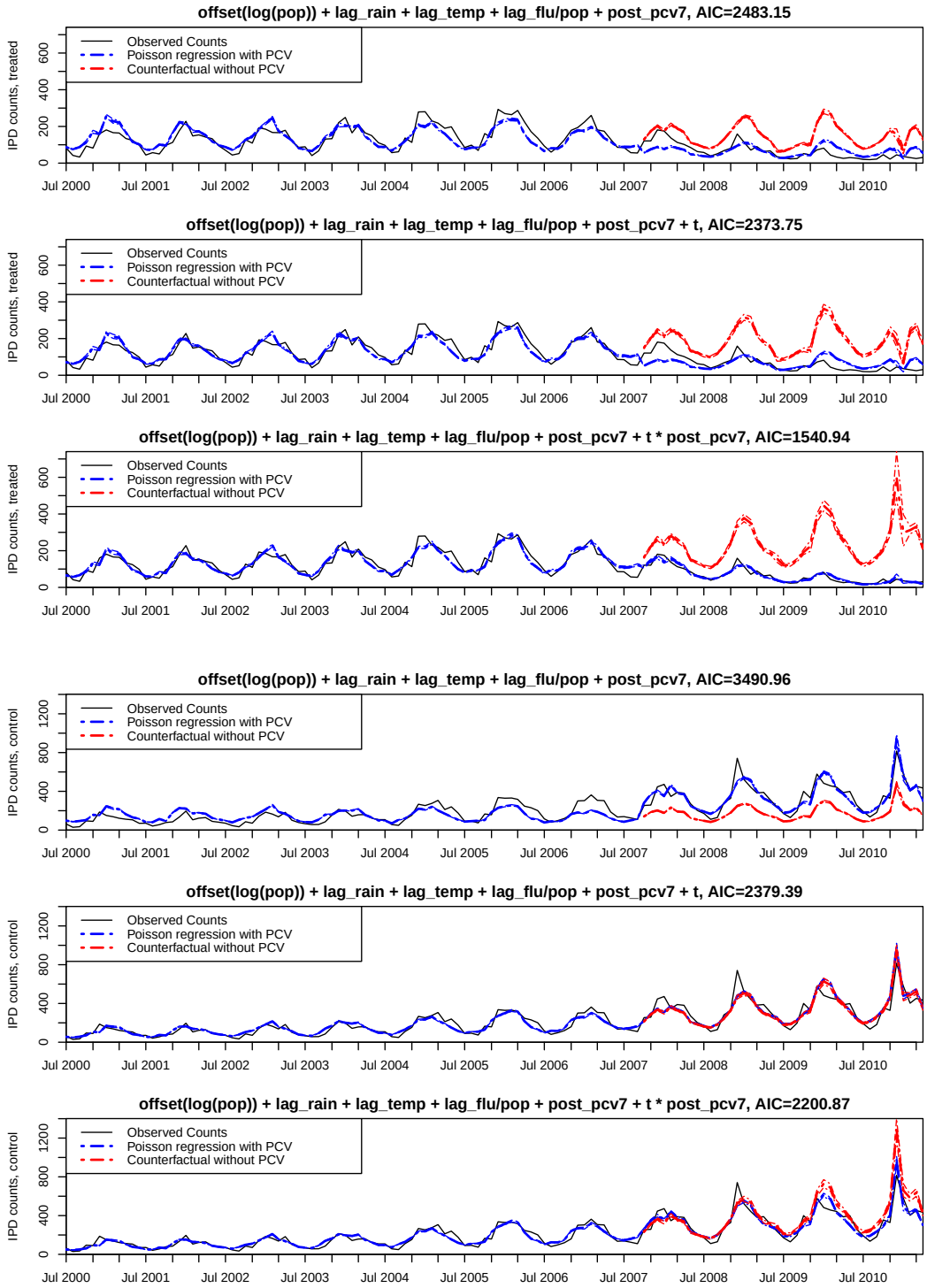


Fig. 6.6 Top three panels: model B, fitted PCV7-IPD counts. Bottom three panels: model C, fitted non-PCV7-IPD counts, based on three ITS models

.

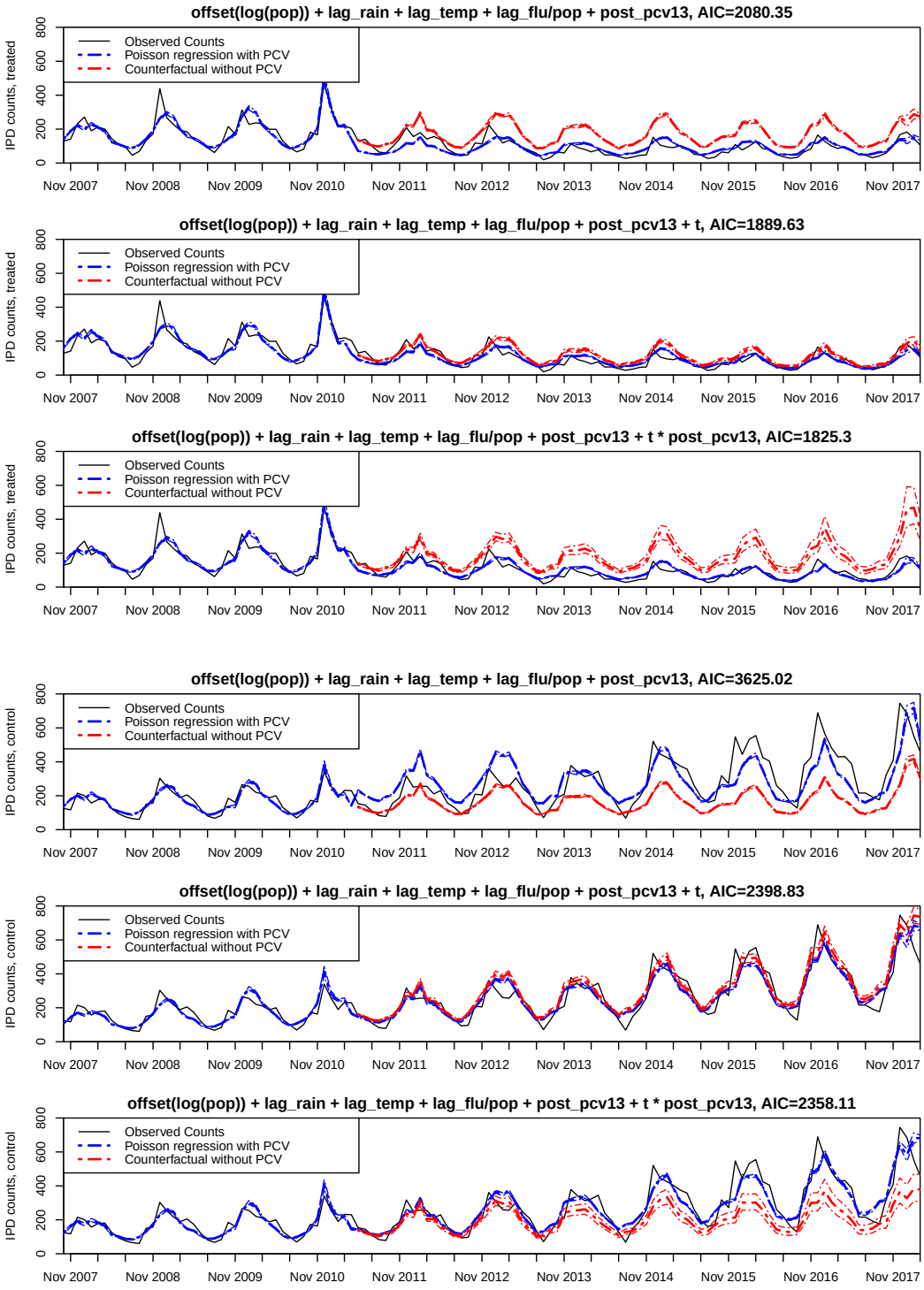


Fig. 6.7 Top three panels: model D, fitted IPD-PCV13 counts. Bottom three panels: model E, fitted IPD-NVT counts, based on three ITS models

.

(CI -11.4%,108.2%) (bottom three panels of Figure 6.7).

6.3.4 CIM

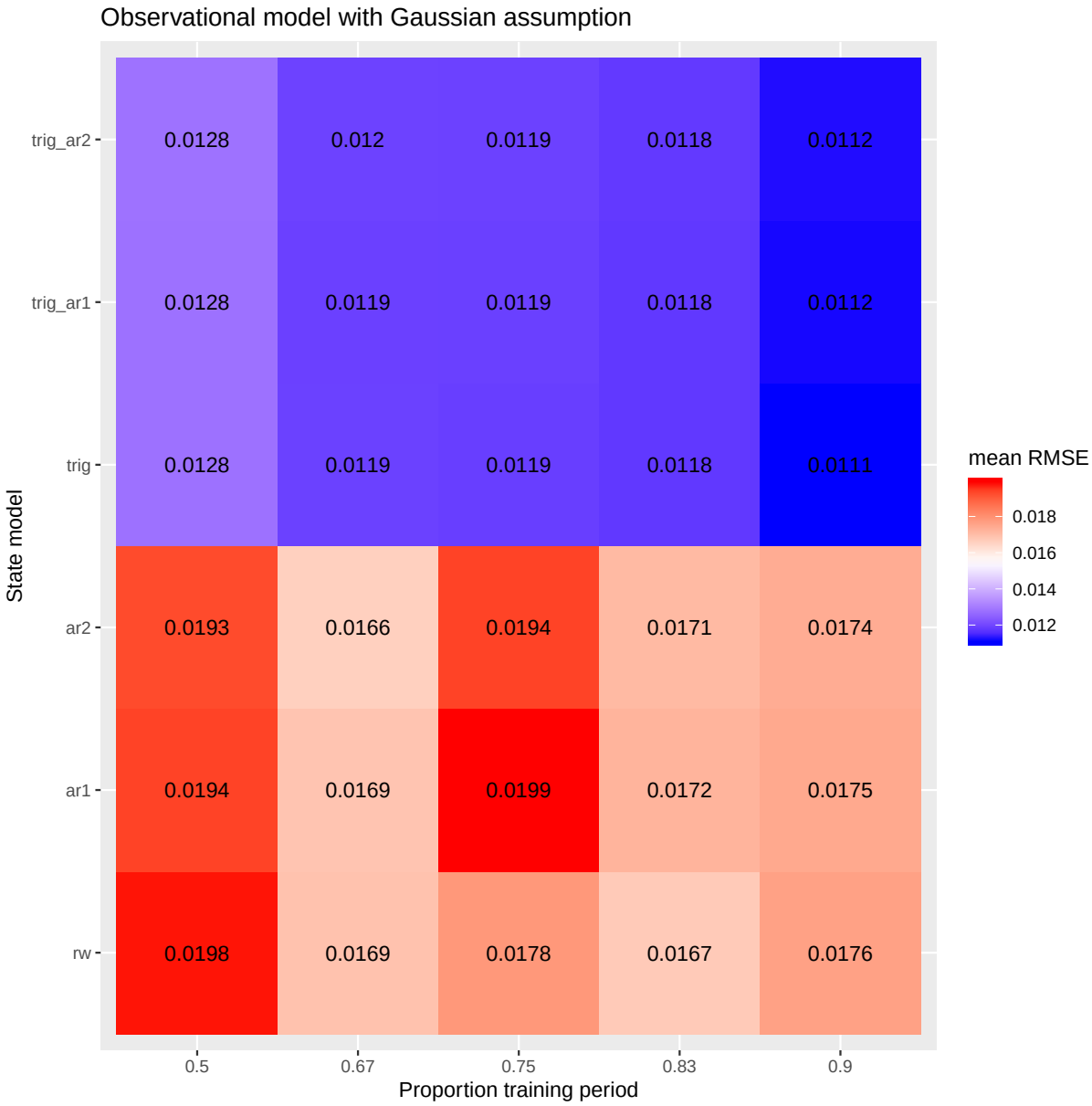


Fig. 6.8 Heatmap of MSPE for different state models and training sets

.

We then model the overall IPD counts using the CIM regression described in section 6.3.2 and, in order to identify the best suited state model, we compare the different fits in

terms of MSPE. Results of the model comparison are presented in Figure 6.8: regardless of the length of the training period (50%, 67%, 75%, 83% or 90%), the smallest MSPE are identified for the models including a trigonometric component, with negligible differences if the trend evolves as a random walk or according to an autoregressive term. Hence, we choose the parsimonious option of a trigonometric component with random walk.

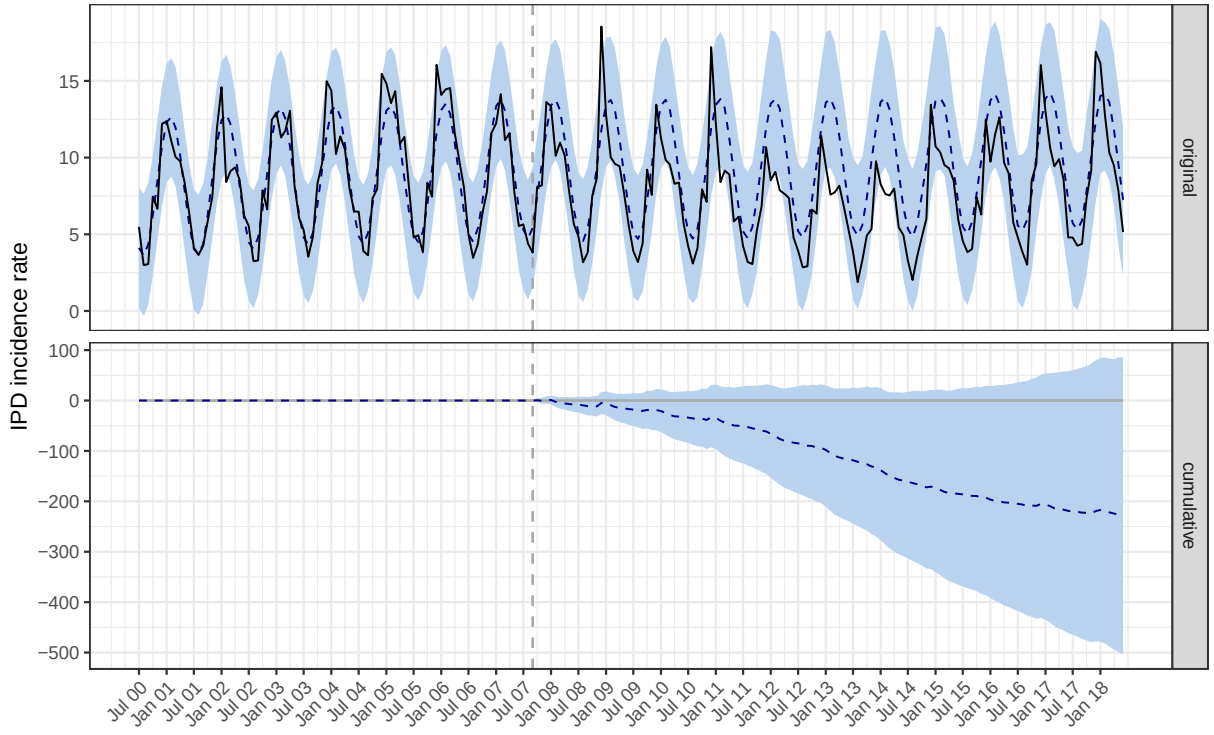


Fig. 6.9 Model A: impact of PCV introduction on the overall IPD incidence rate

.

The estimated impact of PCV introduction on overall IPD (model A) is presented in Figure 6.9. For all graphical representations of the CIM results, the full black line indicates observed incidence while the blue one presents the counterfactual and its credible intervals. Results are plotted in two panels, the top one picturing observed and counterfactual incidence over time, whereas the bottom one shows the discrepancy between them, cumulated over time.

Model A (Figure 6.9) shows that the IPD incidence rate is lower than the estimated counterfactual in the post-vaccine period, but not significantly so (-18.60%, 95% credible interval (CrI) -40.90%, +6.00%). Figure 6.10 summarises the posterior probability for each control time series to be included in the model thanks to the spike-and-slab prior: the trend for

the observed *Haemophilus Influenzae* rate matches the IPD rates best in the pre-intervention period, with posterior probability over 90%. Other control time series are only included with up to 15% posterior probability.

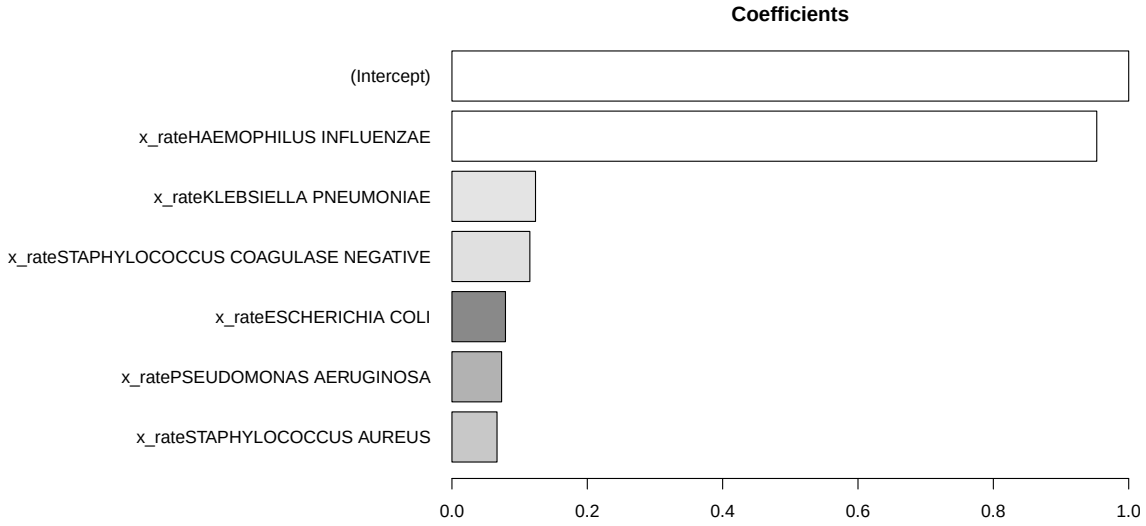


Fig. 6.10 Posterior probability of inclusion for each control time series

We perform the same analysis (model A) across age groups to disentangle whether vaccine introduction has impacted them to a different extent. Results are summarised in Table 6.2: as expected, we find the impact of vaccine to be largest in the children younger than 5 years of age, i.e. the group that did receive vaccination (-49.50%, 95% CrI -64.80%, -40.40%). Reduction in IPD incidence is smaller in non-vaccinated children (5-14), but still significantly bigger than zero. In other groups, instead, the indirect effect due to reduced transmission of S.Pneumoniae after vaccinating children (herd immunity) did not lead to a significant reduction. Age-specific plots are reported in Figures 6.11, 6.12 and 6.13: from the cumulative panel in each plot we can notice how the effect of PCV introduction had immediate effect in children younger than 5, whereas a tendency towards decreasing trends only appeared after a few years in other age groups.

CIM analysis by serotype

When analysing IPD time series by PCV group, we first model the impact of PCV7 introduction on incidence of PCV7-IPD and nonPCV7-IPD (models B and C).

	IRR	IRR_lb	IRR_ub	% change	% lb	% ub
all_age	0.81	0.58	1.48	-18.60	-40.90	6.00
0-4	0.50	0.24	0.92	-49.50	-64.80	-40.40
5-14	0.56	-1.65	3.09	-43.70	-55.80	-31.40
15-44	0.86	-0.07	2.16	-13.50	-38.70	6.00
45-64	1.03	0.42	2.06	2.70	-23.50	22.00
65+	0.90	0.62	1.89	-10.30	-36.10	19.50

Table 6.2 Model A: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV introduction on IPD incidence for different population subgroups.

As shown in Figure 6.14, we find the decline in disease due to pcv7 serotypes to be immediate and substantial in the population overall (model B, top panel), however such a decrease is paired with an increase in nonPCV7-IPD incidence (model C, bottom panel). Table 6.3 summarises PCV7-IPD decrease: -63.9% (CrI -81.0%, -47.6%) overall, with significant effects in all age groups. On the other hand, Table 6.4 presents the magnitude of serotype replacement: a 36.9% (CrI 15.0%, 65.8%) increase is estimated overall, however all age-specific measures fail to show significance, due to the small observed incidence. Age-specific plots are listed in Appendix B, Figures B.8, B.9, B.10, B.11 and B.12.

	IRR	IRR_lb	IRR_ub	% change	% lb	% ub
all_age	0.36	0.25	0.67	-63.90	-81.00	-47.60
0-4	0.36	0.34	0.44	-63.50	-72.00	-53.30
5-14	0.68	-5.38	7.09	-32.50	-66.50	-3.10
15-44	0.70	0.65	0.89	-29.90	-42.40	-17.70
45-64	0.84	0.77	0.95	-15.90	-28.90	-8.80
65+	0.91	0.86	0.98	-9.00	-15.20	-2.30

Table 6.3 Model B: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV7 introduction on IPD-PCV7 serotypes across population subgroups

Secondly, we assess the impact of PCV13 introduction on PCV13-IPD (model D) and NVT-IPD (model E) incidence rates. Results are shown in Figure 6.15. Disease due to pcv13 serotypes shows a sharp decrease (model D, top panel) in the overall population, and once again such a decrease is paired with an increase in NVT-IPD incidence (model E, bottom panel). Table 6.5 summarises impact of PCV13 introduction across age groups: -65.6% (CrI -85.6%, -42.2%) overall, with large and significant impact in all age group. Finally, from Table 6.6 we identify an estimated 31.8% (CrI 6.7%, 63.5%) increase in NVT-IPD incidence overall, however age-specific estimates have substantial uncertainty, and all the

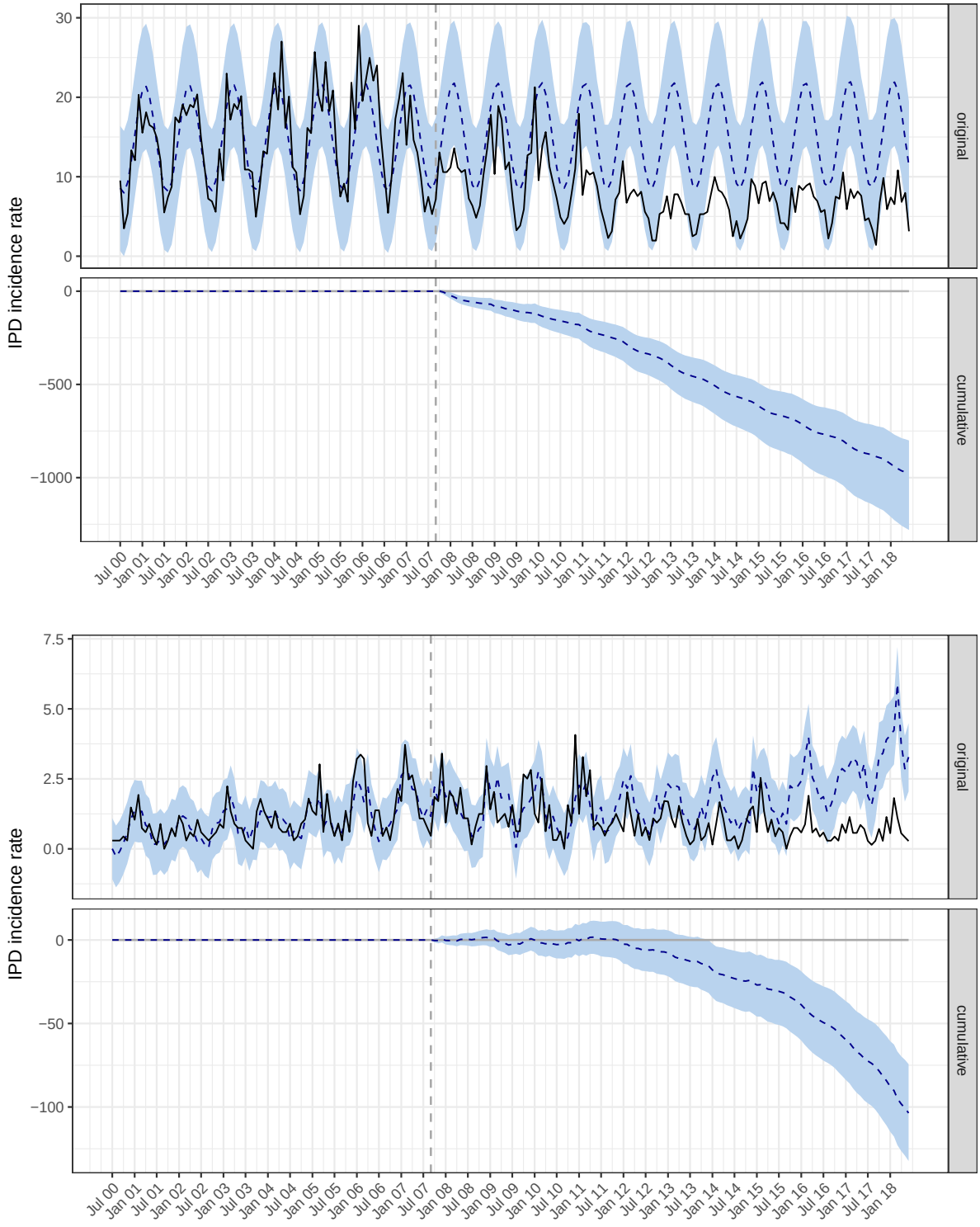


Fig. 6.11 Model A: impact of PCV introduction on the overall IPD incidence rate in children younger than 5 (top panel) and aged 5-14 (bottom panel).

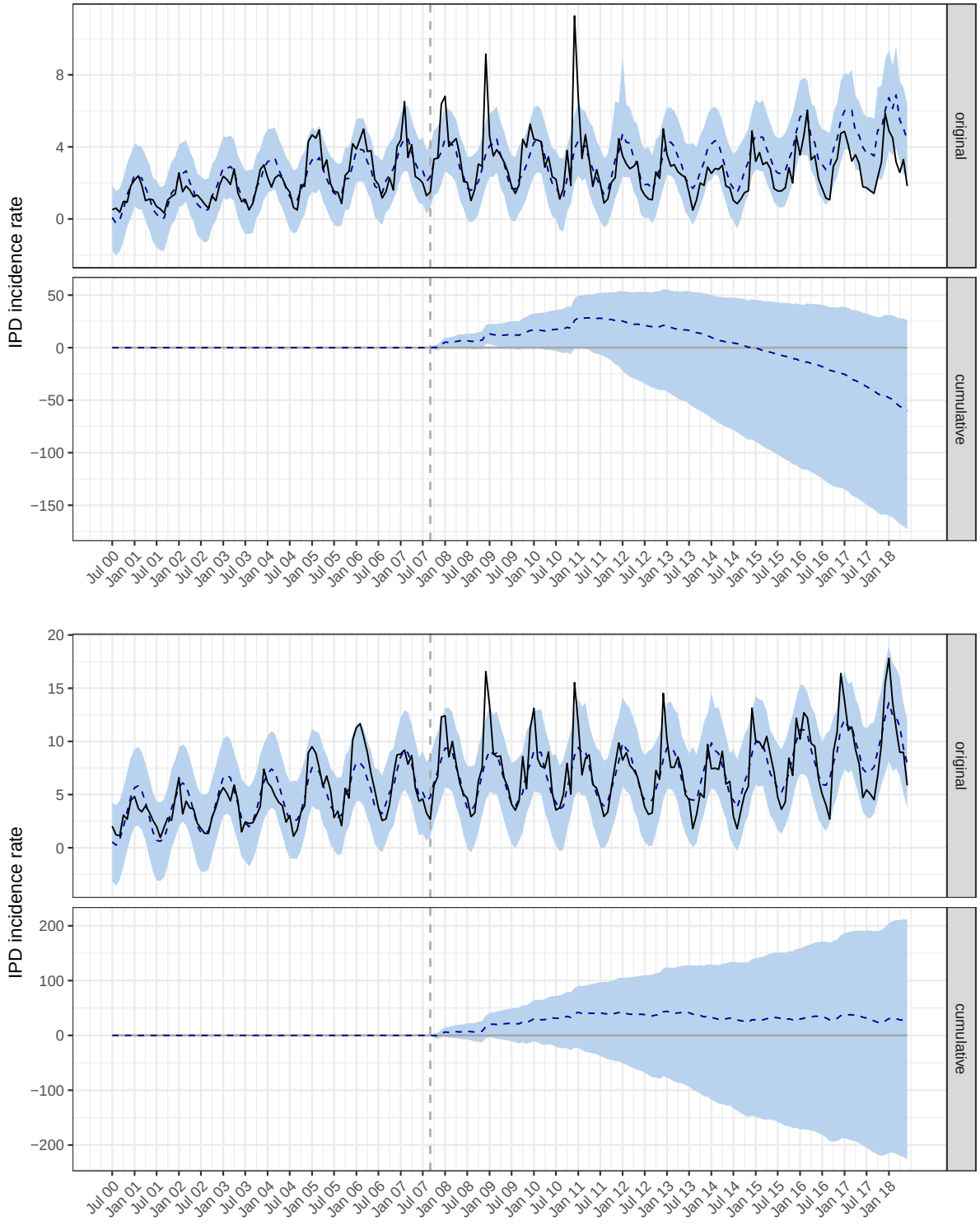


Fig. 6.12 Model A: impact of PCV introduction on the overall IPD incidence rate in adults aged 15-44 (top panel) and 45-64 (bottom panel).

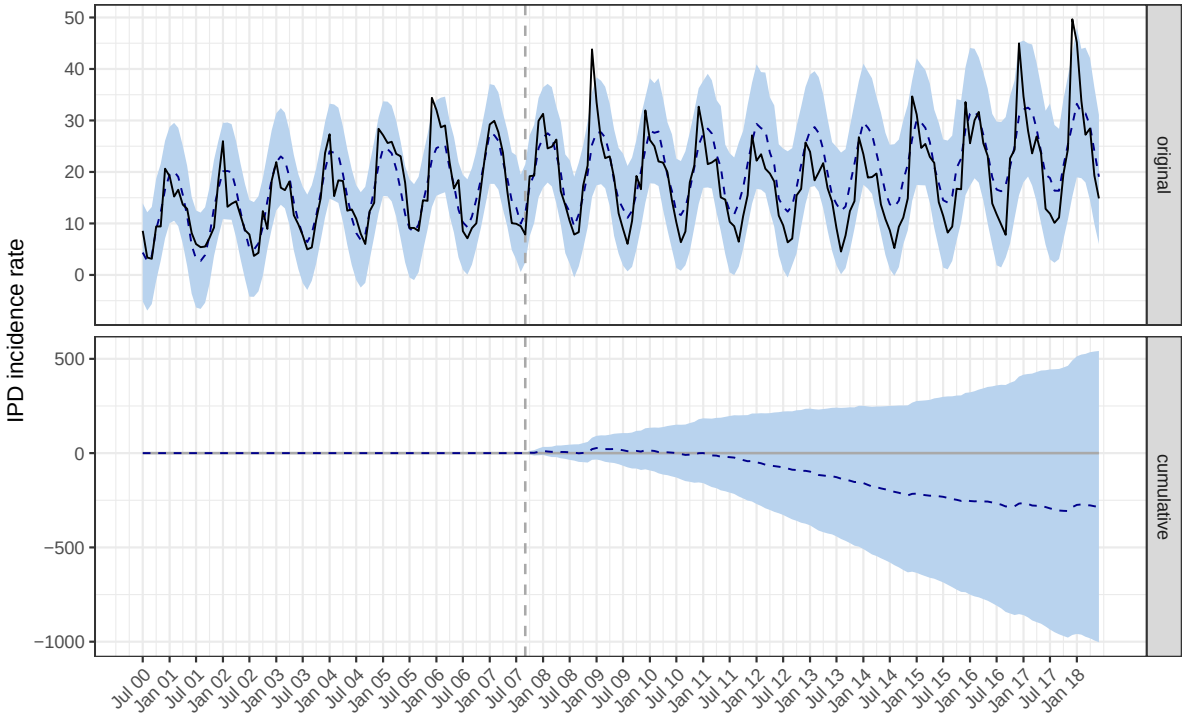


Fig. 6.13 Model A: impact of PCV introduction on the overall IPD incidence rate in the elderly (65+).

corresponding CrIs do include 1. Age-specific plots are listed in Appendix B, Figures B.13, B.14, B.15, B.16 and B.17.

6.4 Discussion

In view of the considerable burden of disease caused by *S. pneumoniae* infection, especially in developing countries, precise evidence on vaccine effectiveness and serotype replacement are needed to make decision on vaccine policies. The impact of pneumococcal vaccines on IPD incidence has been questioned worldwide, however accurate estimates of averted disease burden can be difficult to obtain. Measured IPD incidence exhibits complex temporal dynamics under the effect of changing disease incidence, testing and reporting. In England, an increase in the rate of blood culture sampling over time has been speculated [99]. Further, even before PCV introduction, the distribution of major serotypes responsible for IPD showed temporal variations, likely related to changes in the population immunity. Finally,

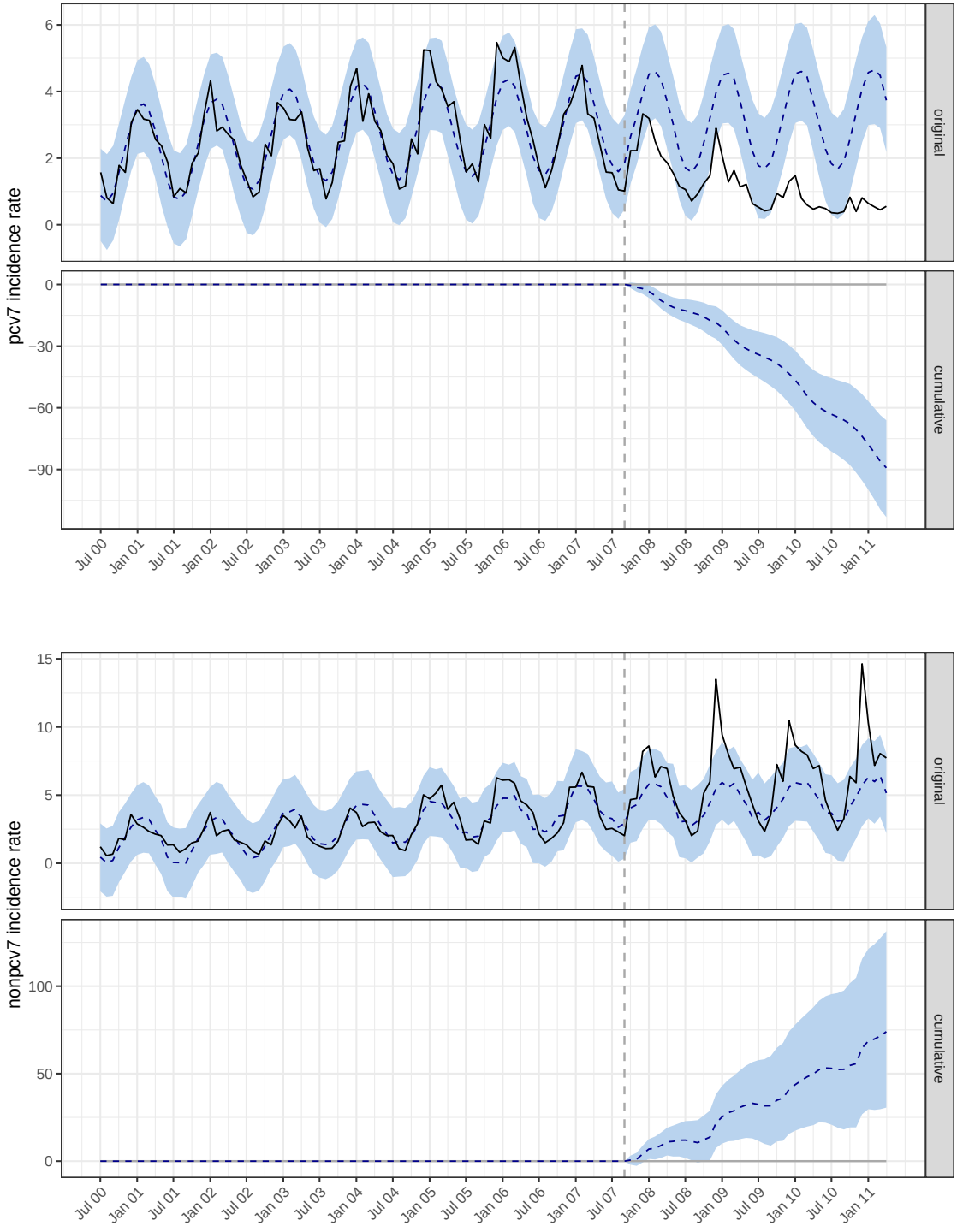


Fig. 6.14 Top panel: model B, impact of PCV introduction on the PCV7-IPD incidence rate. Bottom panel: model C, impact of PCV7 introduction on the nonPCV7-IPD incidence rates.

	IRR	IRR_lb	IRR_ub	% change	% lb	% ub
all_age	1.37	0.85	2.93	36.90	15.40	66.20
0-4	0.78	0.41	1.81	-22.20	-39.40	-3.90
5-14	0.94	-3.52	6.21	-6.30	-26.00	13.90
15-44	0.85	-0.65	3.29	-14.80	-34.90	10.10
45-64	0.82	-0.03	2.60	-18.10	-55.50	14.40
65+	1.04	-2.83	5.91	4.00	-33.90	36.50

Table 6.4 Model C: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV7 introduction on non-PCV7 serotypes across population subgroups

	IRR	IRR_lb	IRR_ub	% change	% lb	% ub
all_age	0.34	0.24	0.67	-65.60	-85.60	-42.20
0-4	0.50	0.46	0.76	-49.80	-63.80	-32.20
5-14	0.53	-1.00	2.90	-46.90	-66.10	-29.10
15-44	0.64	0.59	0.95	-36.50	-57.40	-12.70
45-64	0.81	0.63	1.57	-19.30	-52.80	4.70
65+	0.72	0.64	0.98	-27.60	-46.20	-8.10

Table 6.5 Model D: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV13 introduction on PCV13 serotypes across population subgroups

the proportion of isolates that were fully serotyped also increased over time [208].

In this study, we provide empirical evidence of the effectiveness of a nationwide pneumococcal infant vaccination programme, which led to a 18.6% (CrI -6.0%, 40.9%) reduction in IPD incidence in the population overall, being as high as 49.50% (CrI 40.4%, 64.8%) in children younger than 5 years in England. Such estimates are adjusted for changes in testing and reporting trends, however they are confounded by serotype replacement: the significant impact of the vaccine on the number of infections due to vaccine-targeted serotypes (-63.9%, CrI -81.0%, -47.6% for PCV7-IPD, -65.6%, CrI -85.6%, -42.2% for PCV13-IPD) was compensated by an increase in non-vaccine serotypes. In particular, a 36.5% increase (CrI 15.0%, 65.8%) was estimated following PCV7 introduction, and an additional 31.8% (CrI 6.7%, 63.5%) following PCV13.

Addressing the vaccine effects on the population overall ignores the difference between the direct effects on vaccinated children and indirect effects due to herd immunity: we successfully estimated age-specific effects on vaccine-targeted serotypes, however bigger uncertainty did not allow to make conclusions about different magnitudes of serotype replacement across groups, except for the increase in NVT in children younger than 5, +58.8%

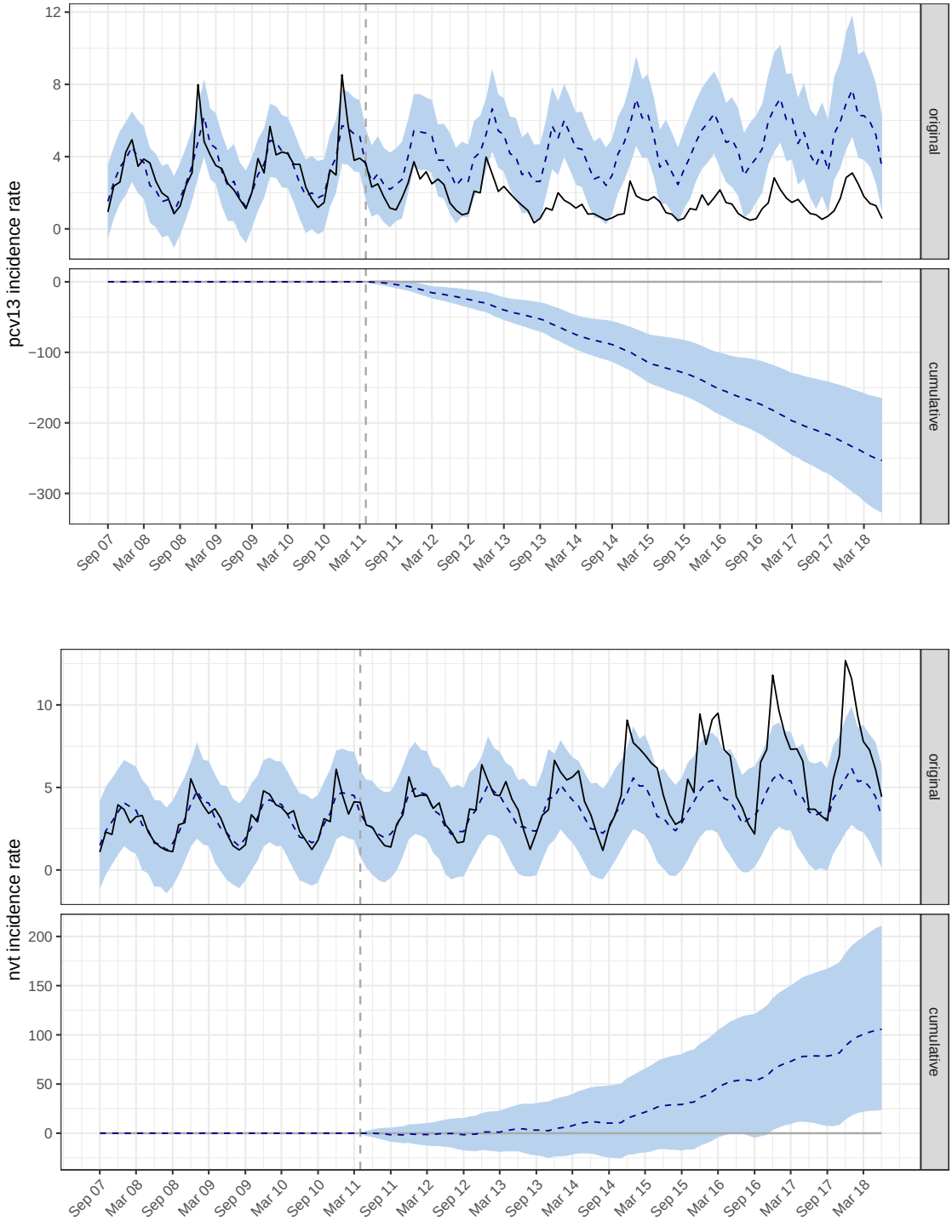


Fig. 6.15 Top panel: model D, impact of PCV13 introduction on the PCV13-IPD incidence rate. Bottom panel: model E, impact of PCV13 introduction on the NVT-IPD incidence rates.

	IRR	IRR_lb	IRR_ub	% change	% lb	% ub
all_age	1.32	-1.00	4.20	31.80	6.70	63.50
0-4	1.59	-6.88	8.65	58.80	24.40	105.00
5-14	1.17	-9.47	10.38	16.90	-19.70	49.10
15-44	1.06	0.92	1.52	6.10	-17.40	21.00
45-64	1.25	-0.96	4.30	25.20	-12.40	49.50
65+	0.88	0.77	1.05	-12.10	-31.50	2.30

Table 6.6 Model E: IRR and relative effects (% change), with 95% CrIs, for the impact of PCV13 introduction on NV serotypes across population subgroups

(CrI 24.4%, 105.0%).

Compared to previous studies modelling the overall IPD incidence rate in England [226, 113, 142], our work shows how diversely the serotype-specific dynamics have been unwrapping across seasons over the last 18 epidemic years. Few studies had previously quantified the magnitude of serotype replacement, but there has been a lack of consistency in the way incidence has been modelled and impact of intervention evaluated across different countries, making generalisations around results difficult. For the sake of comparison with previous work, we initially propose an ITS analysis, with a detailed model to estimate monthly IPD counts, accurately describing seasonality with observed influenza and climate variability, and adjusting for population increases. We also performed a sensitivity analysis to assess robustness of our assumption of one year lag from introduction of vaccination to significant population-wide effects. Despite making full use of the longitudinal nature of the data and accounting for pre-intervention trends through regression modelling, the arbitrary choice of impact model limits robustness of our results. Moreover, due to the considerable length of the study period and to changes to disease surveillance during such period, we acknowledge that an ITS analysis is not the most appropriate research design to account for time-dependent confounding.

In the second part of this work we propose a BSTS, following the CIM. A controlled design proves more useful in evaluating the public health impact of PCV introduction over a long time period, as it allows for simultaneous comparison of the pneumococcal incidence with temporally related infections due to other pathogens, instead of a comparison with previous epidemic years. Moreover, the proposed model incorporates a data-driven approach to select control time series, making the analyst blind and limiting the problem of choosing the impact model arbitrarily.

Both methods lend support to the positive PCVs effects after their introduction in 2006 and 2010, but only the CIM is helpful in discarding the hypothesis that serotype replacement has importantly eroded the benefit of PCV introduction. Since the two classes of models are operating under different assumptions, no direct comparison is possible, however, the CIM methodological framework is more robust, as it does not rely on a pre-defined model choice. Further, sensitivity analyses showed that our results are not sensitive to the choice of priors (results not shown), hence we recommend the use of the CIM methodological framework to evaluate the impact of public health response to constantly changing pneumococcal epidemiology.

Our statistical analyses have a number of limitations. First, we analyse the PCV programme implemented in England and look at data from the past two decades, hence conclusions might only be applicable to other countries with similar pneumococcal dynamics, where no other interventions tackling respiratory pathogens has been introduced. However, the CIM method can be usefully employed using country-specific control time series, to estimate the relevant intervention impact. Second, the choice of pathogens to be used as control time series was made with UK surveillance experts, and it could be questioned when considering a different surveillance system. Nonetheless, the flexibility of the modelling framework, allowing for data-driven selection of the most relevant controls, makes the proposed strategy extremely adaptable. Third, the use of microbiologically confirmed cases does not fully capture the burden of pneumococcal disease, as only invasive disease confirmed by culture from sterile sites is identified. A recent case-only study found increasing trends in hospital pneumonias due to NVT [170]. Further work might be needed to assess the impact of PCV introduction on hospital admissions for respiratory illness, adjusting for time-varying confounding. Thorrington et al. [205]'s study is a first step in that direction, but uses geometric mean to combine control time series instead of a more flexible Bayesian variable selection.

Our findings suggest that the PCV programme is worthwhile in reducing IPD burden at the current rate of serotype replacement, however concerns about a more important serotype replacement in the long term are worrisome.

The variability of serotypes emerging in different countries remains a major challenge in selecting the most beneficial serotypes to add to licensed PCVs. Further work is needed to bridge serotype-specific information from carriage and invasive disease surveillance to genomic studies in order to provide a greater understanding of the post-vaccine adaptation of *Streptococcus pneumoniae* and offer new solutions for future immunisation strategies [98]. Future work might be also needed to investigate whether, reducing carriage of vaccine-type

pneumococci, PCVs are creating an ecological niche favoring colonization by alternative respiratory pathogens such as *Staphylococcus aureus*, *Haemophilus influenzae*, and *Moraxella catarrhalis*.

Chapter 7

Estimated excess all-cause mortality in England during the COVID-19 pandemic

7.1 Introduction

SARS-COV-2 is the newly emerged virus responsible for the current pandemic of COVID-19 disease. It was first detected in early December 2019 in Wuhan, China, and it has quickly spread worldwide. As of June 6th 2020, 6,577,161 cases and 390,748 fatalities associated with COVID-19 disease have been reported globally, however the real magnitude of the SARS-COV-2 burden is still unknown.

Such a burden extends beyond people contracting disease or dying after being infected with SARS-COV-2. Changes in healthcare activity to tackle COVID-19 and interventions implemented by governments during this pandemic are also having an impact on the population's health and mortality. Estimates of excess all-cause mortality can help quantifying the toll of this pandemic. This is not only important retrospectively: sequential quantification of excess mortality is crucial to describing how the epidemic is evolving and to informing decisions on public health strategies for the months to come.

To derive an estimate of such excess deaths, PHE has been using regression models along the lines of the methods described in section 3.2.1: the Euromomo algorithm and a threshold-based outbreak detection method through a Poisson regression model. The Euromomo algorithm [146], adopted by the European Centre for Disease Control, is an approach to quantification of excess mortality due to influenza during the winter period across European countries. This method is based on fitting a Serfling model to weekly counts

of deaths observed in spring and autumn over the past 5 years. This allows estimation of baseline mortality during a period not affected by winter-specific factors such as extreme temperature and influenza circulation through the estimation of a sine wave. This estimated trigonometric function is then used to derive, through extrapolation, the baseline mortality during winter, and excess mortality is estimated as a difference between observed deaths and such a baseline. The second approach uses a Poisson regression [54]: weekly baseline counts are assumed to be independently distributed with mean μ_t, modelled log-linearly as a function of a linear time trend (adjusting for changes in reporting trends) and of death counts in the past 5 years. An overdispersion parameter is estimated when low counts are observed. The 2 and 3 standard deviation prediction intervals are computed, and daily excess is considered to be significant whenever observed counts exceed the upper bound of the interval.

As already discussed in section 3.2.1, these methods suffer from a number of limitations, the most important being that they assume independence among observations. Alternatively, as discussed in section 3.2.2, models including a time series component are more suitable to account for temporal dependence. In particular, the dynamic linear models presented in section 5.4.3 describe the outcome dependency over time through an evolution equation. A second important problem concerns multicollinearity: inclusion of a number of highly correlated covariates in standard regression models leads to estimation problems. In a time series setting, the most significant predictors and lags are often selected a priori based on cross-correlation. A Bayesian variable selection strategy allows including many time series as covariates without selecting them a priori: as explained in section 5.4.1, they can be combined into a weighted average, whose weights are derived to maximise model fit to observed data.

Finally, the estimation of baseline mortality during an outbreak can be cast in terms of estimating a counterfactual for the evaluation of an intervention: a model for the observed mortality before the outbreak is formulated, identifying the most suitable predictors and estimating the corresponding coefficients. The counterfactual, i.e. baseline mortality in absence of COVID-19, can then be forecast. The BSTS methods presented in section 5.4.3 estimate the predictive contribution of multiple covariates through a spike-and-slab prior, and the same regression coefficients are then used to forecast the counterfactual in the outbreak period. Furthermore, the Bayesian credible intervals for the estimated cumulative excess burden provide a natural way to identify the start of an outbreak. Estimation of a counterfactual through a dynamic regression method including control time series, as we did in section

6.3.2, seems a natural alternative to the more traditional regression models.

In what follows we estimate the excess all-cause mortality observed during the COVID-19 pandemic period in England up to the current time as the difference between the observed deaths and a counterfactual estimated through dynamical regression models. We also compare all-cause deaths and COVID-lab-confirmed deaths across population strata, to describe how discrepancies between overall and virus-specific mortality differ across age groups and regions. Finally, we estimate when the excess mortality began in each subgroup, to test whether COVID-19 had circulated undetected in the population before testing started and pandemic was officially declared.

7.2 Data

7.2.1 All-cause deaths

Data on death registrations until midnight on each day in England are provided daily by the General Registry Office (GRO). The daily extract include information about age in years, date of death, death of registration, gender and region. Region is identified according to the district registering the death, which might differ from place of residence for the deceased. As there is a lag between the event of a death and its registration, these data suffer from a reporting delay: counts observed in the recent days are an under-report of the total number of deaths occurred in any of those day. A delay-adjustment procedure through re-weighting of the observed counts, where weights are estimated based on the reporting patterns of the past 5 years, is implemented. The proportion of cases registered at a lag of k weeks (k=0,1,2,3,4, . . . , 52) is estimated through a regression model stratified by age and region to obtain weights for delay correction. Such weights are then applied to death counts for the corresponding age, region and interval from death to the day of the last registrations.

For the scope of the current analysis, daily counts for the epidemic years 2015-2020 are considered, where each epidemic year is defined as starting on the 1st July. The last observed day for the current epidemic year is 29th May 2020. Mortality rates per 100,000 residents are computed for England overall, by gender, by age group (0-24, 25-44, 45-64, 75-84, 85+) and by region (East of England, East Midlands, London, North East, North West, South East, South West, West Midlands, Yorkshire and the Humber).

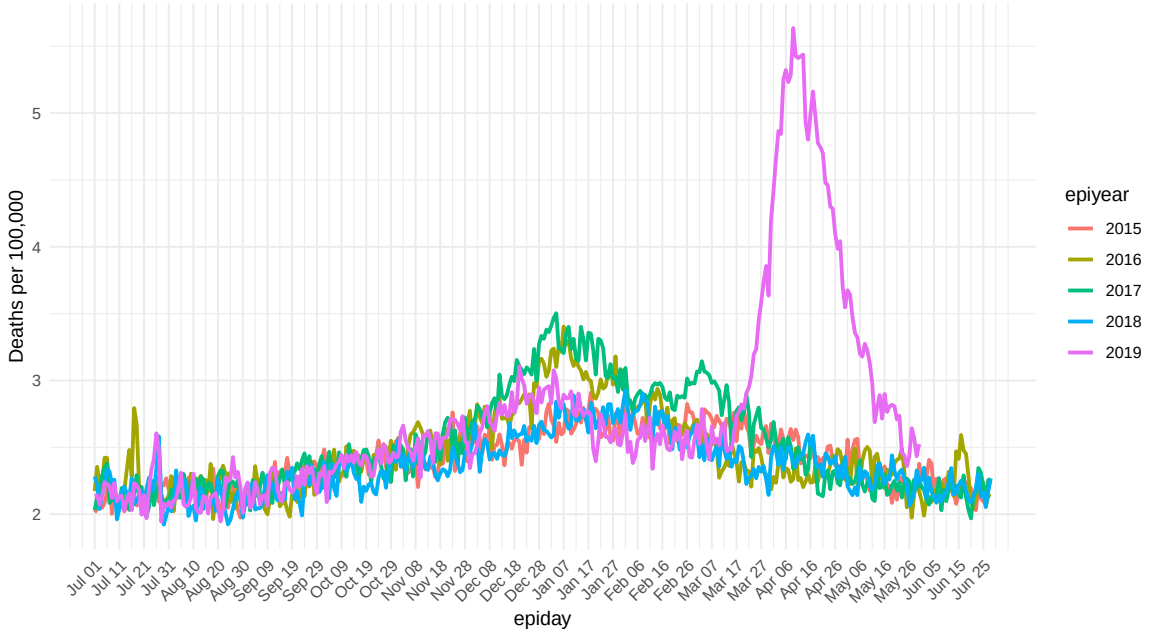


Fig. 7.1 Observed mortality rate in the past 5 epidemic years in England

All-cause mortality rates per 100,000 residents over the study period are presented in Figure 7.1, overlapping the epidemic years for comparison: while the general trends in mortality followed similar patterns across seasons, with only small discrepancies in the amplitude of winter peaks, the epidemic year starting on the 1st July 2019 is characterised by a considerably lower mortality in the months of January and February, followed by an unprecedented spike in mortality in the months of March, April and May 2020, during the first wave of the COVID-19 pandemic. The mortality rate has been increasing steadily until the 8th April, and constantly decreasing from there.

When looking at population subgroups, we see that such an excess of deaths compared to the previous years is equally observed in males and females (Figure 7.2), even though excess mortality in females seems to have a flatter peak and a longer duration. Further, when comparing mortality rates across regions (Figure 7.3), it is evident that the excess emerged in London first, but importantly affected all regions of England. Finally, Figure 7.4 shows the excess deaths across age groups: except for people under 25 years of age, where observed mortality was extremely low at any time, a clear excess in correspondence of the COVID-19 pandemic is identified for all the other age groups.

Fig. 7.2 Observed mortality rate in the past 5 epidemic years in England by gender

7.2.2 COVID-lab-confirmed deaths

COVID-lab-confirmed deaths are provided by PHE and include deaths in those with a laboratory confirmed infection. A delay adjustment is performed along the lines of the one described for all-cause deaths, and rates per 100,000 residents are presented for the same groups as above. Plots are shown in the results section (7.4). We set the start of the COVID-19 pandemic in England to the 2$^{\text{th}}$ March 2020, the day of the first COVID-lab-confirmed death.

7.3 Analysis strategy

We propose a BSTS regression to model Y_t, daily mortality rates of the 2019-20 epidemic year, up to $t < t_1$, where $t_1 = 2^{\text{nd}}$ March 2020. A counterfactual for daily baseline mortality rates $Y_t^{(0)}$ where $t \geq t_1$, i.e. mortality that would have been observed in absence of the COVID-19 epidemic from that date to present, is then forecast from such a model. Excess mortality is

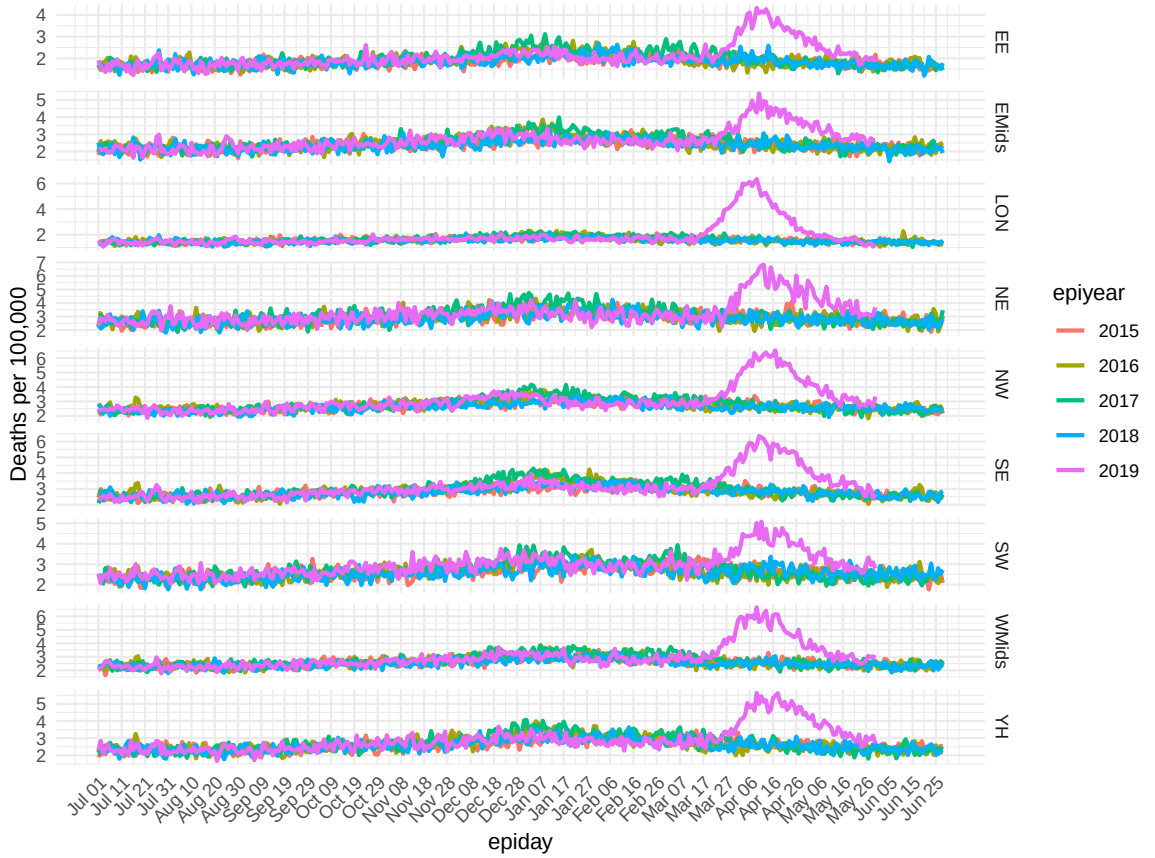


Fig. 7.3 Observed mortality rate in the past 5 epidemic years in England by region

derived as the difference between the observed $Y_t^{(1)}$ and the estimated counterfactual $\hat{Y}_t^{(0)}$.

Our BSTS model assumes a random walk in the evolution equation and a Gaussian noise both in the evolution and the observation equation (details of model formulation have been presented in section 6.3). In formulae:

$$\begin{aligned} Y_t &= \mu_t + X_t\beta + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \eta_t \end{aligned} \tag{7.1}$$

A number of control time series are included through the matrix X_t, selected with the rationale of being potentially good predictors of mortality in absence of the COVID-19 pandemic, and not being affected by COVID-19 pandemic: daily death rates in the past four epidemic years, daily weather conditions such as temperature and rainfall, and counts of other viral infections, such as influenza and RSV, are considered. Their lagged values up to 7 days are also included, as some covariates might lead to delayed effects. This makes a total

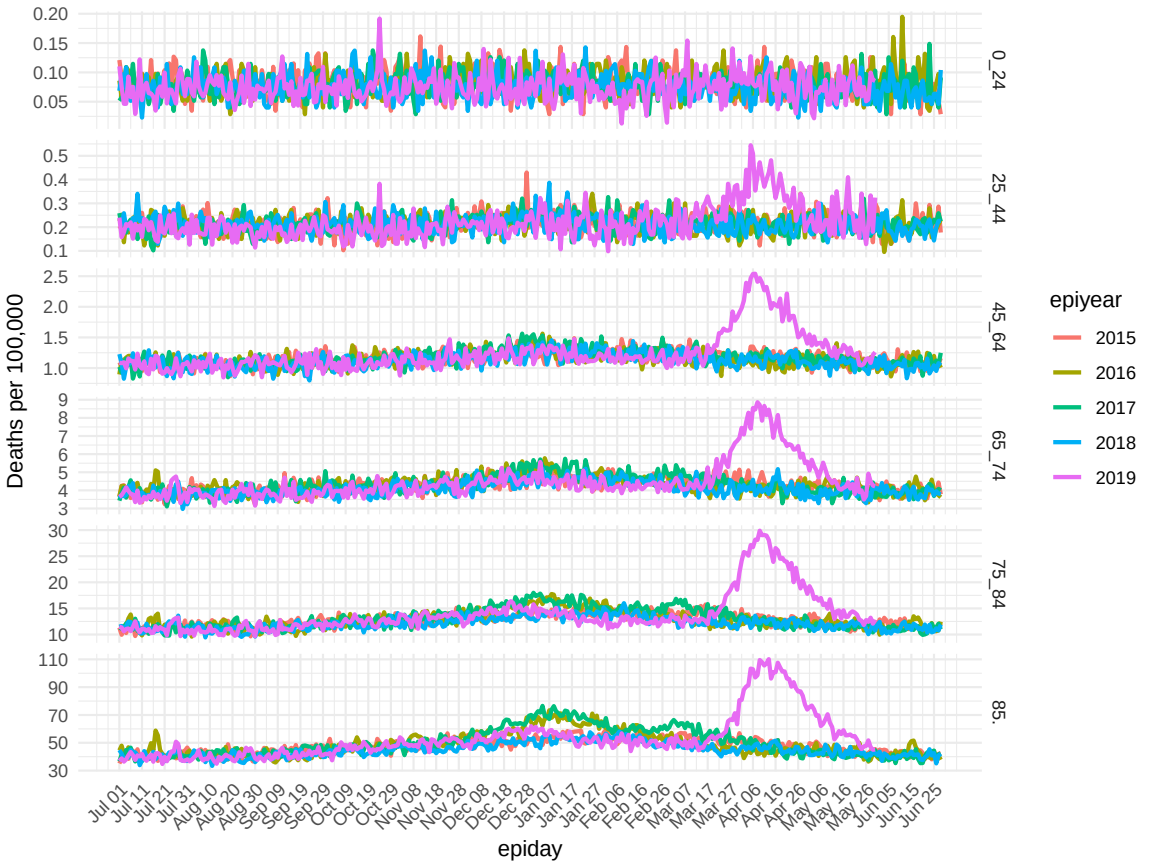


Fig. 7.4 Observed mortality rate in the past 5 epidemic years in England by age group

of 64 predictors.

The control time series are selectively included in the regression component $X_t\beta$ thanks to a spike-and-slab prior on the β coefficients, which allows identification of the most relevant predictors to fit the observed Y_t. We elicit priors by setting the expected model size to be 7, so that the prior probability of each control time series j being included is a Bernoulli($\pi_j = 7/64$). For the slab component, the prior for β coefficients is centered on zero, whereas the Gamma prior for the variance is elicited as a function of the sample variance.

Cumulative excess mortality is quantified as the difference between cumulative observed rates Y_t and the estimated counterfactual $\hat{Y}_t^{(0)}$, for $t > t_1$. The threshold date for outbreak start, i.e. when the excess mortality started being significantly greater than zero, is then identified as the date at which the 95% credible interval for the estimated cumulative excess mortality started being consistently above zero.

This analysis is first performed on the time series of all-cause deaths referring to the entire population, and then replicated for each population subgroup, i.e. by gender, age and region.

7.4 Results

Table 7.1 summarises the excess mortality estimates by age groups and region. An overall cumulative excess of 100.8 (95% CrI 95.8-106.4) deaths per 100,000 residents has been estimated from 2nd March until 29th May 2020 in England. In relative terms, this excess corresponds to an estimated 147% (95% CrI 145%-150%) of the mortality expected during a corresponding spring period, as shown in Table 7.1. This excess represents 164% (95% CrI 156%-173%) of the official COVID-lab-confirmed deaths, that sum up to 34,830 as of 29th May 2020.

	Rate	lb	ub	%baseline	lb	ub	%covid	lb	ub
	100.8	95.8	106.4	147	145	150	164	156	173
F	96.3	90.4	103.2	145	142	148	187	176	201
M	102.8	97.6	108.2	147	145	150	143	136	151
F_65.	434.9	406.6	462.3	143	140	145	178	167	190
M_65.	518.5	490.8	546.5	146	143	148	138	130	145
0_24	0.7	-0.1	1.6	111	98	126	265	-50	620
25_44	7.8	6.5	9.0	143	136	150	264	222	304
45_64	40.0	37.0	42.9	140	137	143	140	130	151
65_74	132.4	122.1	142.7	136	133	139	120	110	129
75_84	473.3	443.6	503.2	142	139	145	127	119	135
85.	1912.4	1559.1	2263.9	144	136	153	189	154	224
EE	69.0	62.8	75.5	141	137	145	108	98	118
EMids	74.6	65.4	83.5	134	130	138	139	122	156
LON	110.9	105.9	116.4	178	175	182	158	150	165
NE	109.7	95.8	124.0	142	137	148	142	124	161
NW	110.8	98.3	121.1	145	140	149	144	128	158
SE	109.8	101.6	118.5	144	141	147	223	206	240
SW	64.2	54.3	74.4	126	122	130	198	168	230
WMids	116.1	107.4	125.1	149	146	153	156	144	168
YH	98.2	87.7	109.3	142	138	147	167	149	186

Table 7.1 Cumulative excess: rate per 100,000, %excess above baseline and %excess above COVID-lab-confirmed, with 95% CrI lower bound (lb) and upper bound (ub).

The posterior probability for each control time series to be included in the model thanks to the spike-and-slab prior is presented in Figure 7.5: among the 64 considered control time series, the model that best predicts the observed mortality rate in the pre-pandemic period more often includes simultaneous counts for mortality rates observed in 2016 and 2017, the influenza circulation lagged by 1 and 2 days, and the amount of precipitation lagged by 7 days.

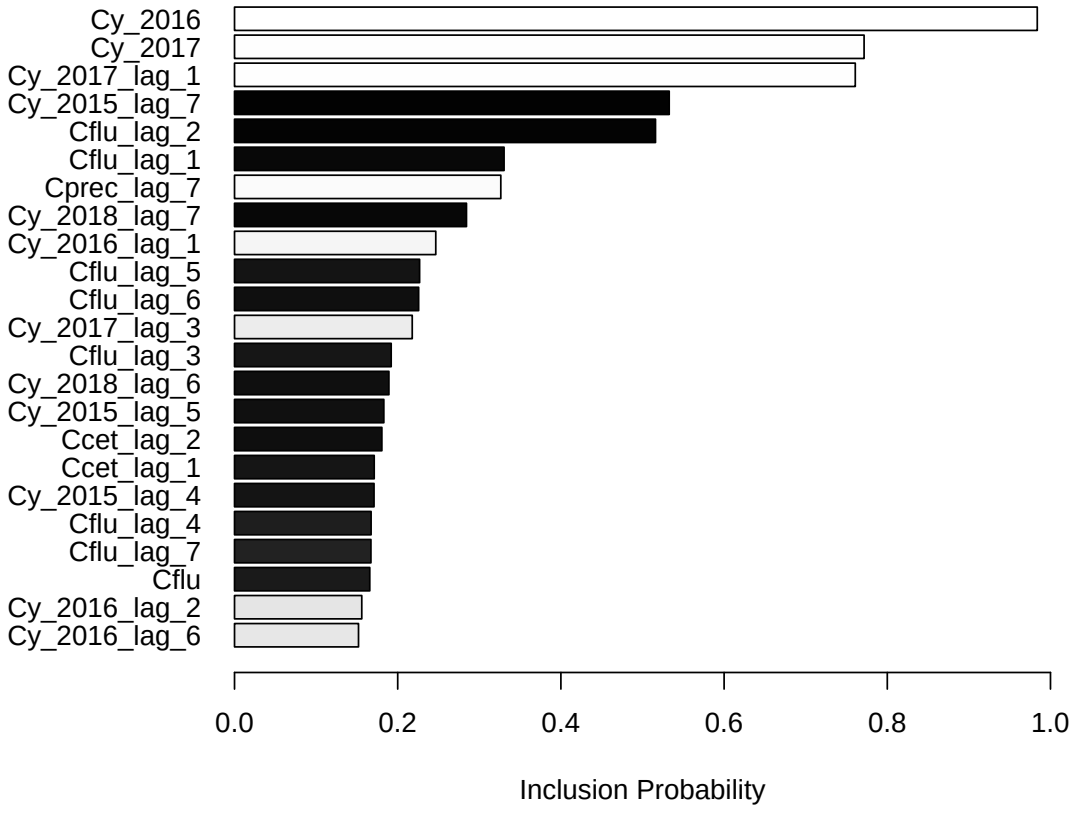


Fig. 7.5 Posterior probability of inclusion for each control time series

7.4.1 Results by gender

The excess all-cause mortality is slightly higher in men compared to women, amounting to 102.8 (95% CrI 97.6-108.2) estimated excess deaths per 100,000 men, and 96.3 (95% CrI 90.4-103.2) per 100,000 women. However, when considering the % excess above the counterfactual, the difference between genders is negligible (results in Table 7.1, column 5). Interestingly, such excess represents 143% (95% CrI 136%-151%) of the COVID-lab-confirmed deaths in men, whereas in women excess all-cause deaths are 187% (95% CrI 176%-201%) of COVID-lab-confirmed deaths. This discrepancy across genders could reflect

a bias in testing, i.e. more men get to hospital and get tested, or could suggest an increased mortality in women for causes other than COVID19.

This result is graphically presented in Figure 7.6, panels 2 and 3, where the blue and red curves are much further apart for women than for men. More in detail, the summary plot for each population subgroup is articulated in three panels: the top panel shows daily observed mortality rates and the estimated counterfactual. The blue area represents the 95% credible intervals around our estimate. In the middle panel the excess death rate is presented in blue, and the COVID-lab-confirmed deaths are superimposed in red; finally, the bottom panel presents the cumulative excess both for all-cause and for COVID-19-confirmed deaths.

7.4.2 Results by age

The estimated excess of deaths is largest for the older age groups, amounting to 473.3 (95% CrI 443.6-503.2) deaths per 100,000 residents aged 75-84, and 1912.4 (95% CrI 1559.1-2263.9) per 100,000 residents 85+ years old (Table 7.1, column 2). It decreases linearly with age, being as low as 0.7 per 100'000 (95% CrI -0.1-1.6) in people younger than 25 (Table 7.1 and figure 7.7). However, these discrepancies across age groups are proportional to the corresponding baseline mortality rates: comparing excess during the COVID epidemic to the counterfactual (Table 7.1, column 5), it is clear that **deaths have been between 36% and 44% higher than expected in all age groups except for the youngest**, with minor differences across groups.

On the other hand, the proportion of excess all-cause deaths with respect to the COVID-lab-confirmed deaths differ greatly across groups: it is smallest in 65-74 and 75-84 years old, being 120% (95% CrI 110%-129%) and 127% (95% CrI 119%-135%) respectively. Blue and red curves are very close to each other in figure 7.7. Instead, it is much larger in 25-44 and 85+ years old, covering 264% (95% CrI 222%-304%) and 189% (95% CrI 154%-224%) of COVID-lab-confirmed deaths (figure 7.7). Confidence intervals are wide due to small numbers, however a potential bias, either with less testing and hospital admissions, or more deaths due to other causes, could be suspected for these age groups.

Finally, important differences can be seen across age groups also in terms of when the excess deaths started (Table 7.2): the age group 25-44 signals an early start (14th March), followed by the 45-64 years old on the 19th March, and reaching 75-84, 85+ and children

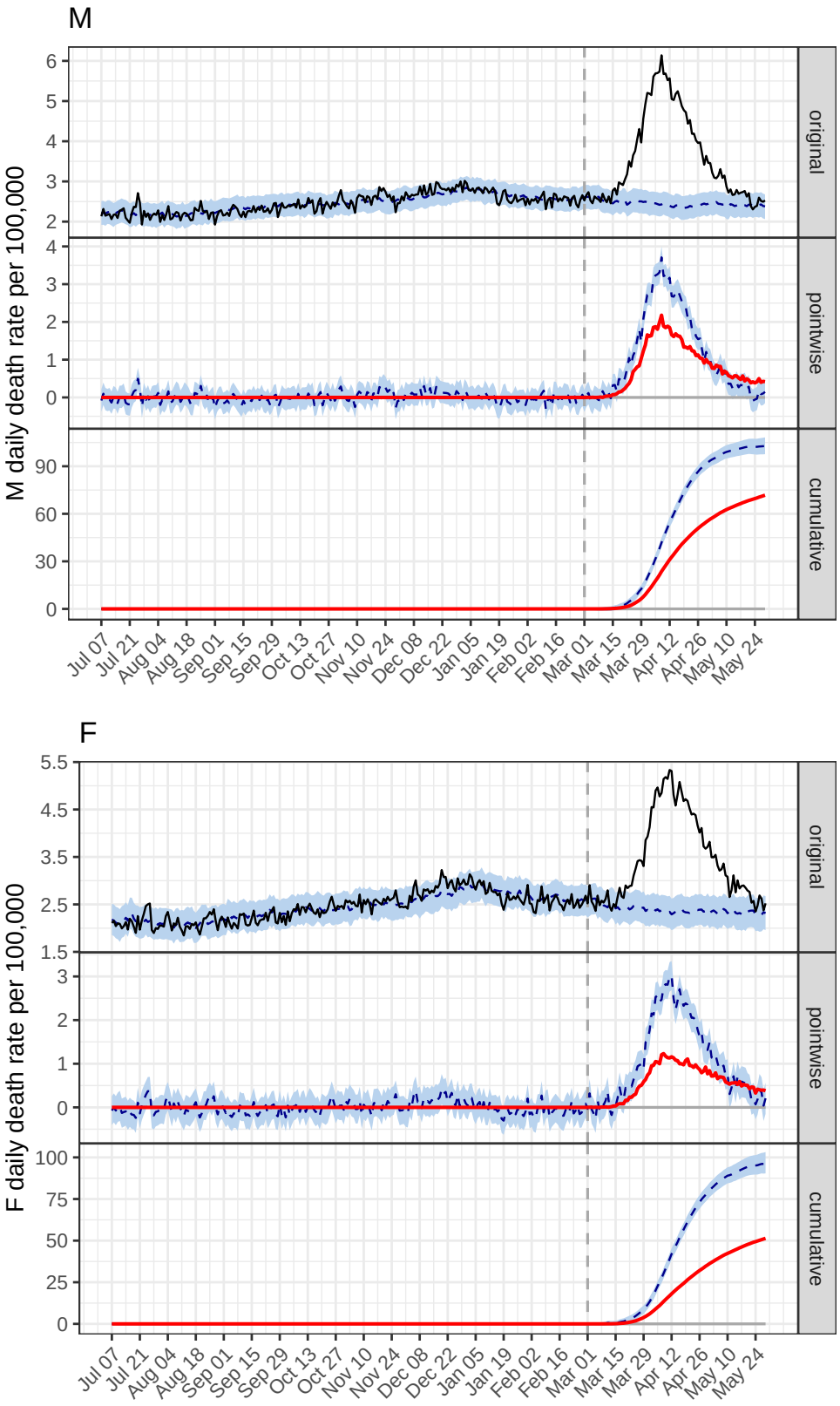


Fig. 7.6 Excess all-cause mortality for men (top panel) and women (bottom one)

last (26th, 25th and 27th March respectively).

7.4.3 Results by region

116.1 (95% CrI 107.4-125.1) extra deaths per 100'000 residents have been estimated for West Midlands, the most affected region, with a rate that is almost two times the one observed in the South West (64.2, 95% CrI 54.3-74.4). Other heavily impacted regions are London, North West and South East, with 110.9 (95% CrI 105.9-116.4), 110.8 (95% CrI 98.3-121.1) and 109.8 (95% CrI 101.6-118.5) extra deaths per 100'000 residents respectively.

However, when comparing the proportion of deaths with respect to seasonal baseline, London emerges with a strikingly high 178% (95% CrI 175%-182%), while all other regions report excesses between 34% and 49%. Finally, the largest excess of all-cause mortality with respect to COVID-lab-confirmed deaths is observed in South East and South West (223%, 95% CrI 206%-240%, and 198%, 95% CrI 168%-230% respectively) while in East of England the excess all-cause deaths almost overlaps with the COVID-lab-confirmed counts (108%, 95% CrI 98%-118%).

Results for England overall and for each region are plotted in Figures 7.8, 7.9 and 7.10. In particular, in Figure 7.8 we present the excess all-cause mortality in England and in London to show how the excess mortality emerged earlier in London compared to England overall. Table 7.3 summarises such a comparison across all regions: West Midlands and London saw early increases (16th and 19th of March respectively, and the lag from the first regions was as long as two weeks for the North East (1st April).

7.4.4 Comparison with Poisson regression

Figure 7.11 shows a comparison over time for the absolute number of excess deaths estimated with the BSTS and the Poisson regression model. Table 7.3 shows the same comparison across population subgroups. Making use of the BSTS modelling framework we estimate 53,476 excess deaths (95% CrI 45,912-61,154) from the 2nd March until the 29th May. Over the same time period, the PHE Poisson regression model estimates an excess of 58,701. When grouping by age and region, PHE point estimates lie within the BSTS credible intervals in all cases except for London. However, we notice they generally are above the posterior median and in the upper tail of the posterior distribution for groups where incidence is higher

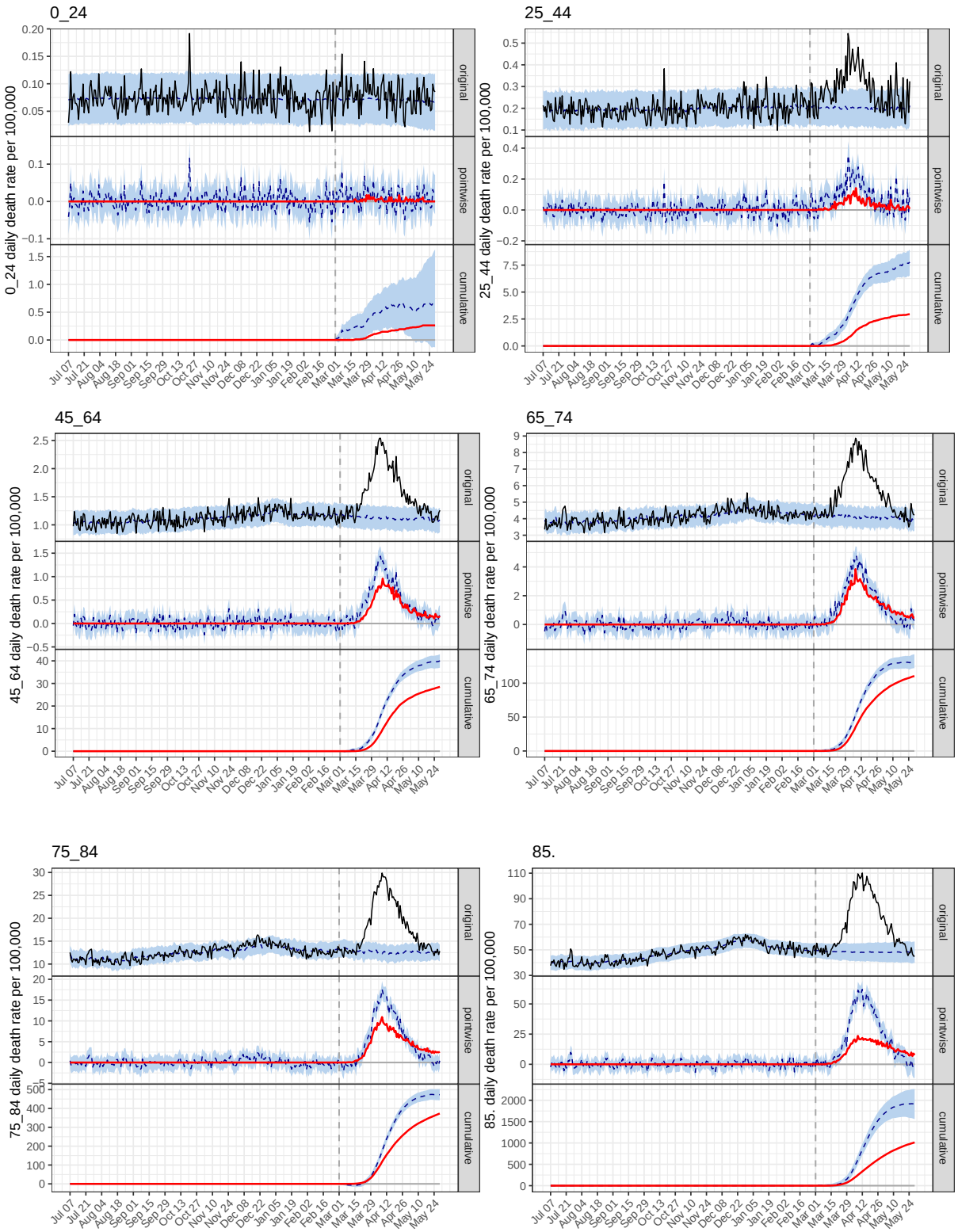


Fig. 7.7 Daily excess in all-cause mortality by age group, per 100’000 residents

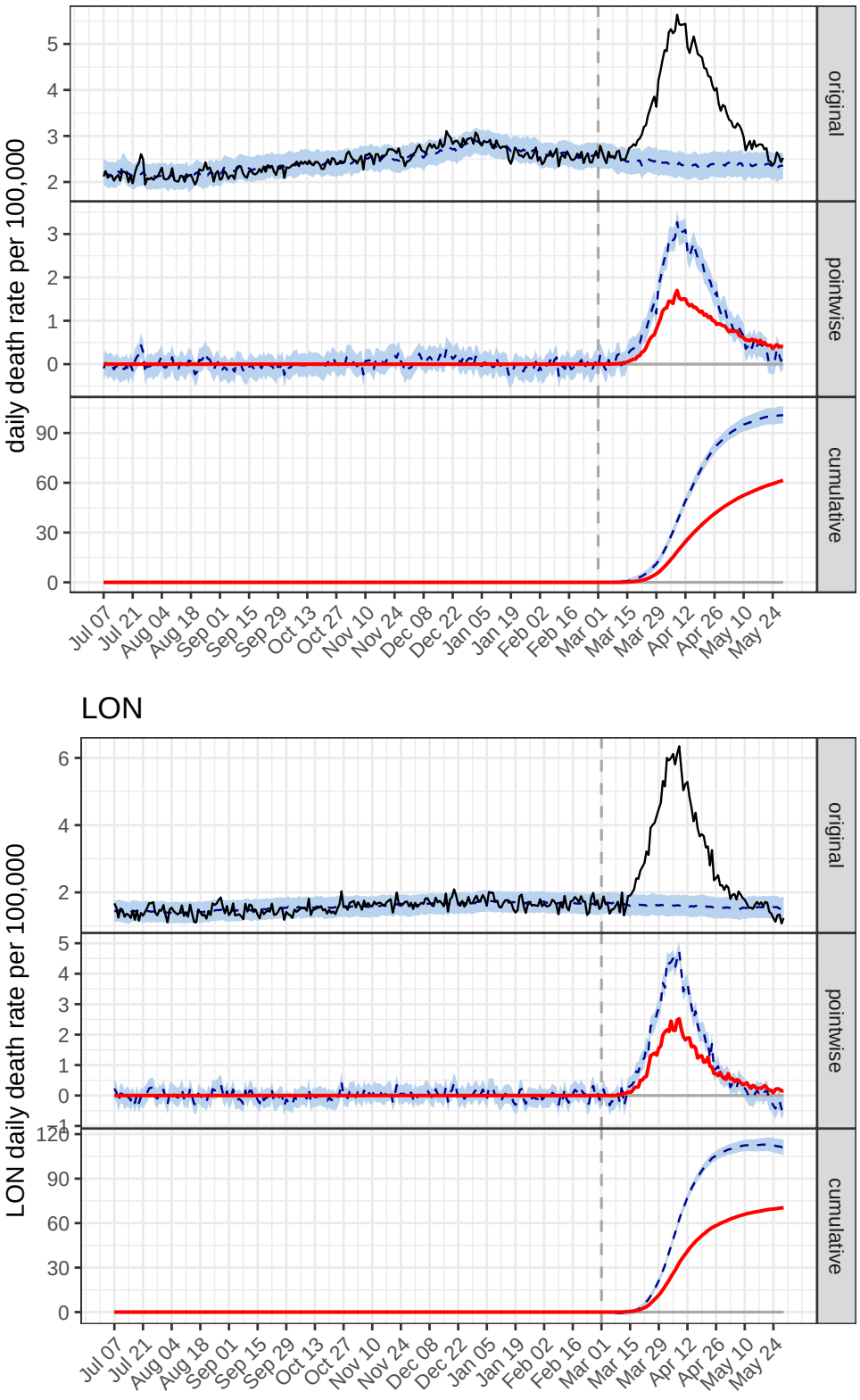


Fig. 7.8 Excess all-cause mortality in England (top panel) and in London only (bottom panel)

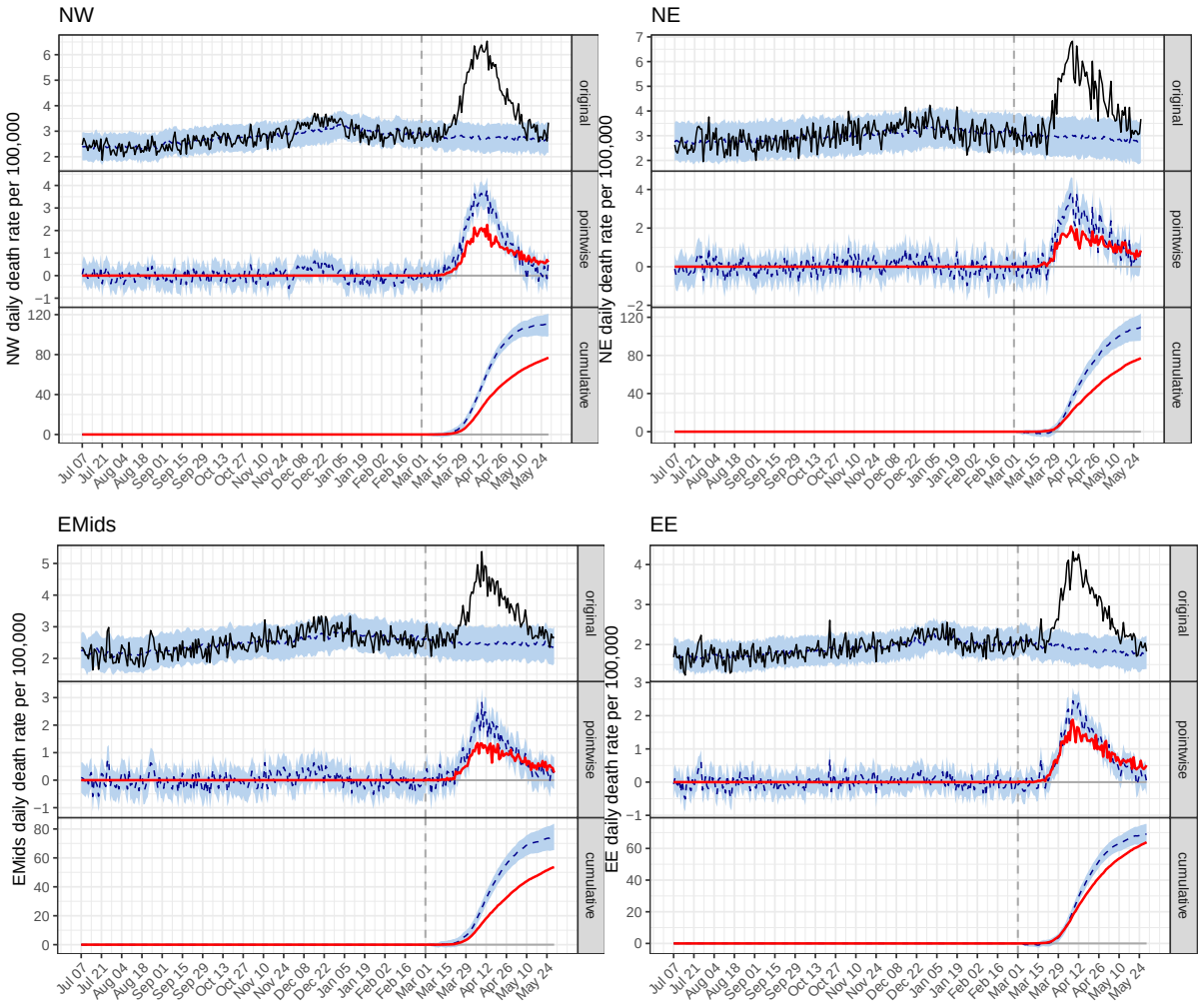


Fig. 7.9 Excess all-cause mortality for North West, North East, East Midlands and East of England

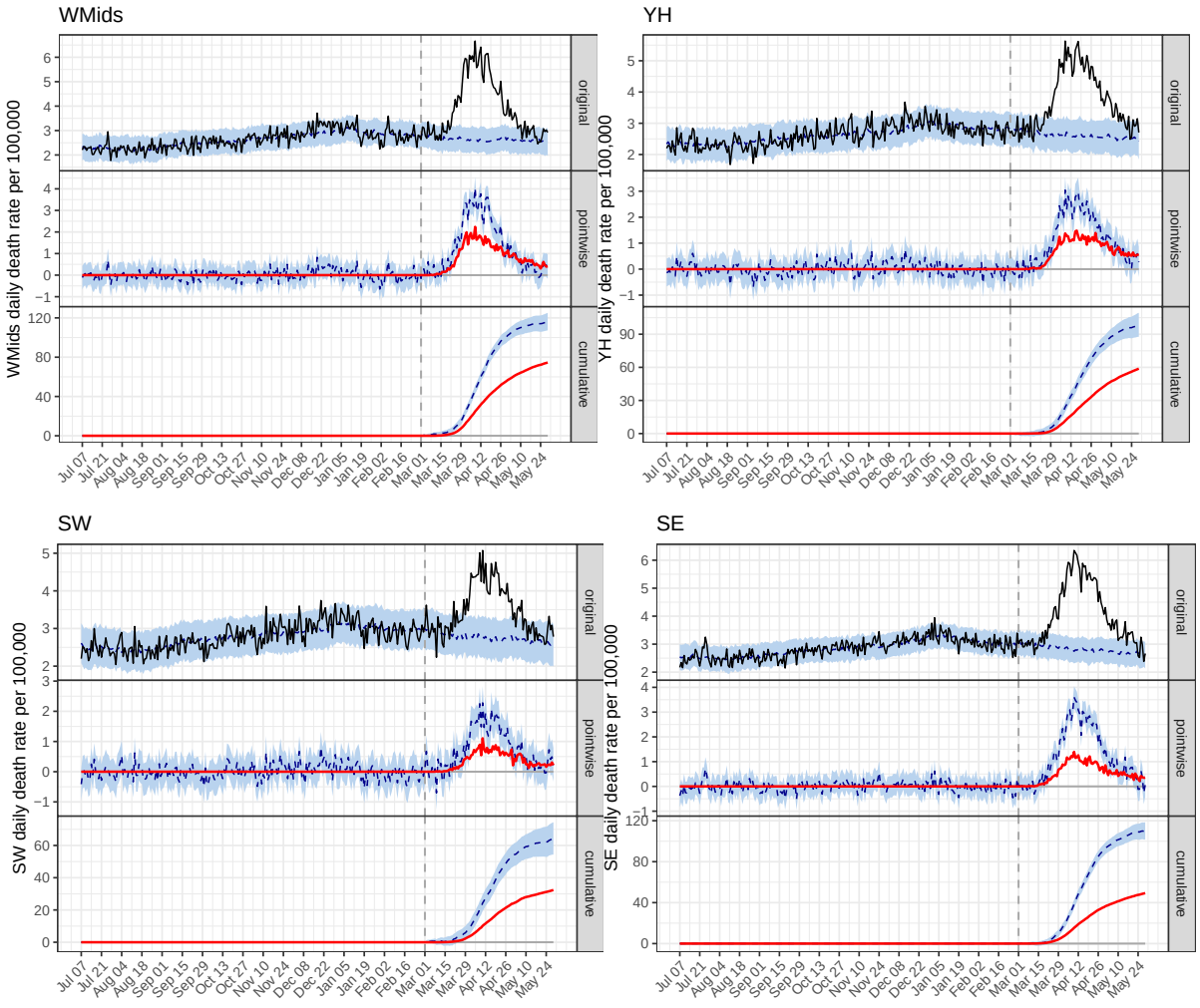


Fig. 7.10 Excess all-cause mortality for West Midlands, Yorkshire and the Humber, South West and South East

	date	N	lb	ub	Rate	lb	ub
	2020-03-18	601	516	687	1.13	1.08	1.20
F	2020-03-20	286	233	338	1.08	1.02	1.16
M	2020-03-19	308	258	357	1.16	1.10	1.22
F_65.	2020-03-24	257	208	307	4.89	4.57	5.19
M_65.	2020-03-22	259	217	302	5.83	5.51	6.14
0_24	2020-03-27	2	-0	3	0.01	-0.00	0.02
25_44	2020-03-14	13	11	15	0.09	0.07	0.10
45_64	2020-03-19	65	60	70	0.45	0.42	0.48
65_74	2020-03-20	83	76	89	1.49	1.37	1.60
75_84	2020-03-26	179	145	212	5.32	4.98	5.65
85.	2020-03-25	256	210	303	21.49	17.52	25.44
EE	2020-03-28	49	45	54	0.78	0.71	0.85
EMids	2020-03-26	40	26	54	0.84	0.74	0.94
LON	2020-03-19	113	108	119	1.25	1.19	1.31
NE	2020-04-01	33	28	38	1.23	1.08	1.39
NW	2020-03-24	87	64	111	1.25	1.11	1.36
SE	2020-03-23	114	102	125	1.23	1.14	1.33
SW	2020-03-26	35	19	50	0.72	0.61	0.84
WMids	2020-03-16	73	54	91	1.30	1.21	1.41
YH	2020-03-26	61	52	69	1.10	0.99	1.23

Table 7.2 First date of excess mortality and average daily excess: number of deaths and rate per 100,000

(e.g. when only grouping by gender, age groups 75-84 and 85+, London, West Midlands).

Beyond point estimates, the two methods differ in terms of quantification of uncertainty: the assumption of independence among observations used in the Poisson does not appropriately account for temporal dependence, leading to overconfident estimates, whereas the BSTS allows propagation of uncertainty, with credible intervals becoming wider the further ahead we forecast.

7.5 Discussion

Using English national registry data, we quantify the magnitude of all-cause excess mortality in England, and compare our estimates with the ones released by PHE. Compared to other methods, we get a potentially more realistic reconstruction of the baseline because we account for time-dependence of observations. Further, we are able to include more controls, and we

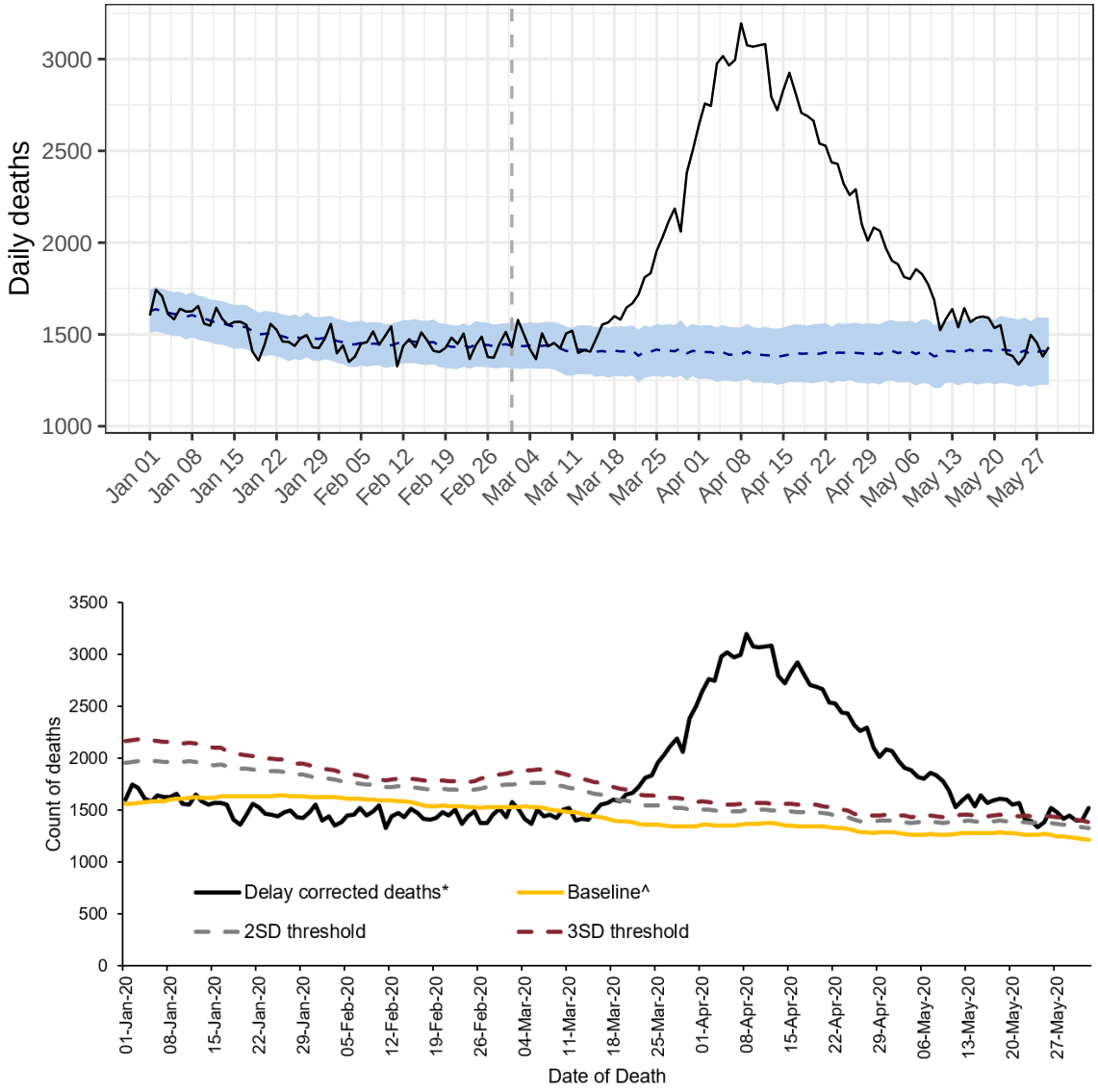


Fig. 7.11 Observed all-cause deaths and estimated counterfactual obtained with the BSTS model (top panel) and with the Poisson regression (bottom panel) for England up to May 29th

	Synthetic controls			Poisson regr
	N	lb	ub	N
	53476	45912	61154	58701
F	25436	20778	30058	28814
M	27423	22959	31762	29867
0_24	136	-11	308	93
25_44	1162	979	1341	1108
45_64	5805	5366	6218	5612
65_74	7363	6754	7935	7326
75_84	15931	12908	18892	18333
85.	22823	18646	26966	26211
EE	4380	3985	4781	4341
EMids	3603	2341	4812	4028
LON	10091	9622	10604	10880
NE	2932	2502	3353	3147
NW	7760	5673	9894	9067
SE	10151	9059	11100	10047
SW	3151	1699	4452	4312
WMids	6525	4810	8140	7696
YH	5441	4644	6160	5574

Table 7.3 Cumulative excess: number of deaths estimated using synthetic controls (with 95% CrI) and with Poisson regression

naturally obtain uncertainty through the Bayesian posterior distribution of our counterfactual. For age groups with lower mortality rates, where observations are volatile, our estimates naturally reflect the uncertainty due to small numbers.

Our results show an important excess all-cause mortality over the COVID-19 pandemic period, and comparison with the recorded COVID-19 deaths shows that the undetected excess mortality concentrated in the early phases of the epidemic, whereas in recent weeks the two measures match. While it is expected that the majority of excess deaths from COVID-19 in the short term are from the virus itself, further work is needed to rule out the extent to which excess deaths are directly due to COVID-19 infection or to other causes, such as additional pressures on the health and social care system. While the NHS has increased ICU capacity and converted several wards to host COVID-19 patients, lower standards of care for non-COVID-19 emergencies may be characterising hospitals at this stage. For example, a recent study conducted in 54 Italian hospitals by De Rosa et al. [45] found that hospital mortality from heart attack tripled from 4.1% to 13.7% in March 2020 compared to the same period last year, suggesting that hospitals were unable to adequately treat and care for

patients with other conditions.

As of 11th June 2020, the EuroMOMO study of 24 European countries shows that on average the continent is returning to normal levels of mortality, with effectively no excess mortality in France, Spain or Italy. England is still ranked as having "moderate excess" [ECDC], however our estimates show that this is only the case in selected regions and age groups. Moreover, the current estimates only represent a partial toll of the overall burden of the ongoing pandemic and up-to-date estimates will be produced in the upcoming weeks.

An intriguing hypothesis about excess mortality during the COVID-19 pandemic is the harvesting one, i.e. many individuals who died from the virus were frail, and their events were only brought forward by a brief period of time [67]. If this is true, we should expect lower mortality rates in the months to come. However, other factors might affect the long-term risk. Patients may not be attending hospital from fear of contagion, reaching hospital with delays, at increasingly serious conditions and with complications, which make life-saving treatments much less effective. A&E attendances in England were down from 2.2m in May 2019 to 1.3m in May 2020, a drop of 42% [152], and research conducted in early May 2020 showed that 47% of people would feel uncomfortable using their local hospital in the short term if the need arose, three quarters of whom (76%) would be concerned about being exposed to COVID-19 [Ipsos Mori]. The Italian study mentioned above [45] also found a reduction in hospitalizations for heart attacks homogeneously throughout the country, despite the fact that the COVID-19 pandemic hit Northern Italy hardest.

Further changes to healthcare activity implemented to protect patients and to free up NHS resources, such as cancellation or postponement of elective surgeries and other non-urgent treatments, might also have long-term effects. The number of patients admitted for routine treatment in hospital in England was down by 85 per cent, from 280,209 in April 2019 to 41,121 in April 2020 [154]. 60% fewer people with suspected cancer were urgently referred to a specialist in April 2020 compared to the same month in 2019, and the number of first treatments for cancer fell by 21% [153]. The number of people having to wait more than 18 weeks to start treatment rose to 1.13 million, almost double the number in April 2019 (579,403) and the highest number for any calendar month since January 2008 [155]. Such a re-prioritisation of non-urgent care may also result in additional deaths over the future months depending on the length of the delay in treatment and on how such a delay would affect the specific outcomes [133].

Finally, factors unrelated to healthcare access but rather depending on the restrictive measures could have a potential impact on the excess deaths: a decrease in the deaths due to car accidents, work accidents and violent crime is expected, as opposed to an increase in deaths resulting from domestic violence and suicides due to anxiety and depression [82, 16, 147]. The negative impact of this pandemic on the economy, with rising levels of unemployment and deprivation, might also affect mortality in the long term [53].

Chapter 8

Empirical dynamical modelling

8.1 Introduction

Up to this point we have focused on stochastic methods for linear time series. Stochastic models attribute the observed variability of an outcome of interest, pneumococcal disease incidence in our case, to one or multiple exogenous random variables, e.g. influenza and meteorological conditions, plus some random noise. Further, the assumption of linear dependence within and across time series of interest makes them straightforward to interpret and to use for forecasting. Section 3.2.2 presented how, in the univariate situation, linear dependence between Y_t and Y_{t-h} can be quantified by the autocorrelation function (equation 3.4) or by including the autoregressive component Y_{t-h} in the linear predictor of a regression model for Y_t. Similarly, when looking at multiple time series, the cross-correlation function and inclusion of X_t and X_{t-h} as predictors for Y_t estimate linear dependence between time series Y_t and X_t.

Despite testing the relevance of multiple lags, we have struggled in identifying strong evidence to support such a linear relationship. Therefore, in this chapter we question the choice of deeming a linear stochastic model appropriate to describe the underlying system. In the next sections we introduce nonlinear time series (NLTS) theory, we describe its use with infectious disease data, present details for some empirical dynamical models (EDM) and conclude with an application to IPD and influenza time series.

8.1.1 Why a nonlinear time series analysis?

In statistics, the shape of the system which produces observations is generally assumed to be known to some degree: a model is formulated, and then inference is performed to match such a model to observations. In the case of time series analysis, the choice of linear stochastic models is ubiquitous, and the theory underlying the assumption of linear system is expressed by the Wold Decomposition: any stochastic process can be separated into the sum of two processes, a deterministic one that is a linear function of its past values, and a stochastic one that is a linear function of previous values of an uncorrelated random variable [175]. This implies that, when a time series shows an irregular behaviour, univariate linear time series methods can only attribute such an aperiodic behaviour to exogenous variables.

Such a priori insight about the system dynamics might be misleading: dynamical laws governing nature or human activities are seldom linear, and complicated temporal behaviours are exhibited by deterministic systems in the presence of nonlinearity, resulting in a condition sometimes referred to as deterministic chaos [108]. Hence, the apparent randomness observed in one or multiple time series could be the expression of an underlying complex deterministic process rather than stochasticity.

Further, when multiple time series are observed, absence of linear relationships between them might induce to think that they are not related in the underlying system. For example, the Lorenz system was first defined in 1963 by Lorenz [124] using a set of three first-order differential equations. Such a system is an approximation to atmospheric convection : x, y and z are underlying variables that allow describing rates of change of horizontal temperature, vertical temperature and convection in a two-dimensional fluid layer as a function of physical parameters σ, ρ and β.

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y-x) \\ \frac{dy}{dt} &= x(\rho-z)-y \\ \frac{dz}{dt} &= xy-\beta z \end{aligned} \tag{8.1}$$

Despite the deterministic coupling expressed in equations 8.1, projections of this three-dimensional system onto each axis X, Y and Z as a function of time result in three time series which are not linearly correlated to each other, as shown in Figure 8.1: dependence between X and Z is quadratic (panel 2), while no correlation can be identified between Y and Z (panel 3).

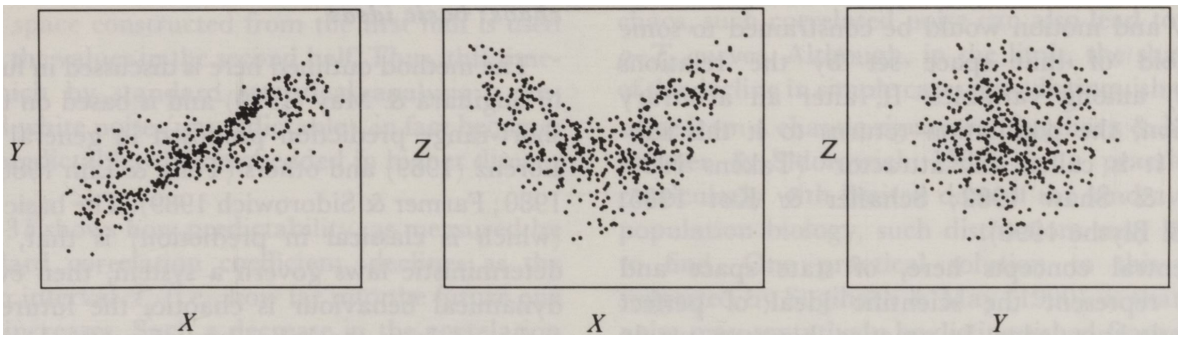


Fig. 8.1 Correlations between all pairs of variables for the Lorenz attractor [198]

Hence, when dealing with one or multiple observed time series that have complex, highly variable or random appearance, instead of aiming to estimate what the current state of a known system is, how noise influences the resulting observations, and what the next observations will be, we could instead aim to learn more about the nature of such a system. For example, it could be of interest to reconstruct how the system evolves over time, or to test whether two variables belong to the same dynamic system and how they relate to each other.

Beyond reconstructing the system for understanding or for forecasting purposes, diagnosing the structure most closely corresponding to reality would be valuable also with the scope of designing an intervention. In particular, ascertaining whether the observed variability is driven by stochastic or deterministic dynamics can be crucial when proposing interventions: if you assume that you are dealing with a stable system except for exogenous shocks, you expect such a system to self-correct, and you only consider implementing policies tackling exogenous factors. On the other hand, if a deterministic dynamic is likely, policies that could prevent future boom and bust cycles should be taken into consideration [97].

The branch of mathematics which analyses the long-term qualitative behavior of dynamical systems is called chaos theory: it aims to reconstruct the unobserved geometry of a dynamical system in order to explore whether the system will settle to a steady state in the long term, what is the possible underlying attractor (i.e. the set of values toward which the system will evolve, which encapsulates its long-term dynamic behaviour), and how this depends on the initial conditions.

A set of empirical tools for the study of complicated time series dynamics, generally referred to as EDM, has been object of extensive research over the past 30 years. These methods, which have been largely used in signal processing and engineering, share a philosophical approach based on inductive reasoning: they process information on real-world data

without making model assumptions explicit. Regularities as well as non-obvious patterns are detected in the observed time series with the scope of extracting information and formulating hypotheses, understanding if the system dynamics are linear or nonlinear, and diagnosing presence of stochasticity [97].

This approach has proven to be fruitful in the understanding of many complex phenomena, nonetheless very few natural systems have actually been found to be low dimensional deterministic [84]. Only in those cases, i.e. when a deterministic model that describes the system dynamics exists, other NLTS methods can assist in finding solutions for those equations [97]. Even though identifying the governing equations of a dynamical system is a convenient way to describe its long-term behavior, in most cases NLTS analysis gets to the point of identifying characteristics of the underlying system without being able to explicitly write down a model for it.

8.1.2 EDM in infectious disease

Historical analysis of ecological time series relied on modelling measurement error and considering key information to be enclosed in the underlying smooth patterns. However, with the emergency of chaos theory suggesting that irregularities can be an equally interesting object of investigation, application of NLTS tools flourished in many settings[198].

In particular, the relative importance of deterministic versus stochastic forces in ecological populations has been object of a long-standing debate, and the discussion focused in particular on environmental versus biological factors: environmental factors were thought to be associated with stochastic fluctuations in population density, and biological ones with deterministic regulation [198].

Within the field of infectious disease, the apparent correspondence between real-world epidemic time series and the chaotic properties of simple epidemiological models have attracted considerable attention [206]: the irregularity in duration and amplitude of cycles of measles epidemics in the pre-vaccination era has been the focus of an extensive search for non-linearity and chaos, with many attempts of reconstructing characteristic properties of the underlying system from a single observed time series [183, 160].

More recently, Deyle et al. [46] used novel EDM tools to investigate environmental drivers of influenza outbreaks. They speculated that the relationship between these quantities

might not be linear as influenza outbreaks are well correlated to the seasonality of temperature and absolute humidity in temperate countries, while seasonality both of climate and outbreaks is much weaker in tropical countries. In order to test for the presence of such a nonlinear relationship they reconstructed the two-dimensional system from country-level epidemic and climatic time series. Following their work, Cobey and Baskerville [28] investigated from simulated data to which extent the periodic, noisy, and transient dynamics of ecological systems could be an obstacle to system reconstruction and any deriving causal inference.

8.2 Methods

8.2.1 Dynamical systems and their geometry

In mathematics, a dynamical system is characterised by three components:

1. a state of the system at any given time, $\mathbf{x}(t)$, i.e. a vector containing the values assumed by the system internal variables at time t
2. an evolution rule $\dot{\mathbf{x}} = v(\mathbf{x})$, i.e. a function that describes what future states follow from the current one, typically formalized into one or multiple differential equations
3. a state space, representing the totality of states the system could visit, e.g. $\mathbb{R}^n$

A geometric representation of the state space can effectively help to visualise a dynamical system. A space of dimension n is considered, where each axis represents one of the n internal variables which uniquely define the system dynamics, and the vectors $\mathbf{x}(t)$ are used as vectors of coordinates, so that each $\mathbf{x}(t)$ is represented by a unique point in the corresponding state space. The system evolution over time can also be represented in the state space by joining the sequence of states visited according to the evolution rule, an object called system trajectory. For example, the Lorenz system trajectory is pictured in the top panel of Figure 8.2.

As anticipated in section 8.1.1, the interest is usually on the long-term behavior of the system, when the system trajectory might converge to the underlying attractor. An attractor can be a point, a curve, a surface, or even a complicated set with a fractal structure. However, initial conditions can affect the long-term dynamics of a system: each trajectory represents the set of states visited when starting from one particular initial condition, and exploring variations entails plotting the trajectories compatible with starting from *any* initial condition. Such a visualisation proves extremely useful to understand qualitative features of the system

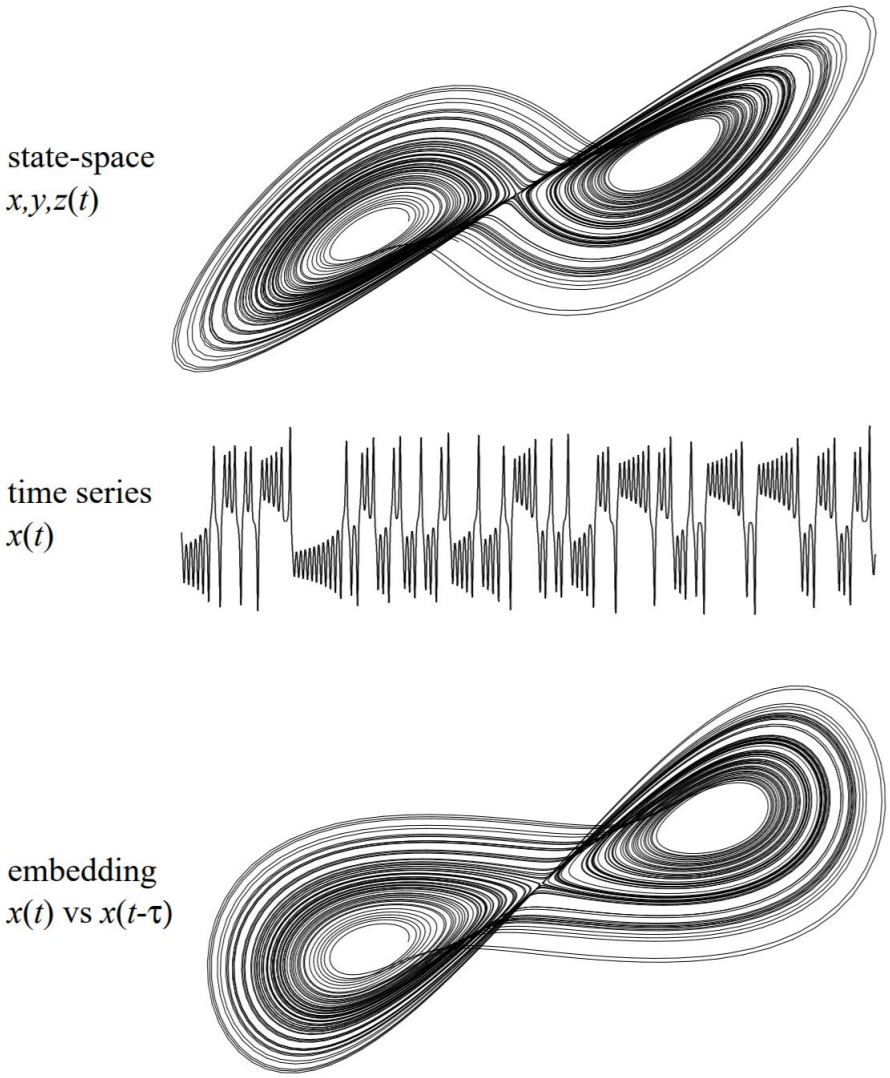


Fig. 8.2 Top: representation of a Lorenz system trajectory in the x, y, z space, where the variables are temperature, pressure gradient and angular velocity. Middle: time series of $x(t)$, discarding any knowledge of y, z, or the governing equations. Bottom: phase-space reconstructed by embedding $x(t)$.

when it is impossible to derive an explicit formula for the solution of its nonlinear equations, elucidating qualities of the system that might not be obvious otherwise [197].

If a deterministic system is defined by ordinary differential equations, its state space is finite-dimensional and the attractor can be represented geometrically. For example, instantaneous states $\mathbf{x}(t)$ of an idealized pendulum are uniquely defined by its angle and angular velocity; its state space, a two-dimensional state plane, is the set of all possible pairs "(angle, velocity)", and the path drawn by the pendulum visits converges to a closed curve, a type of attractor typical of systems presenting periodic oscillations. On the other hand, if the system is three- to n-dimensional, the attractor can be a manifold, as the butterfly-shaped Lorenz attractor in Figure 8.2. However, the number of dimensions can grow unexpectedly, making state spaces difficult to use: in more complicated cases, as in partial differential equations and delay differential equations, the state space can be infinite-dimensional [197].

8.2.2 Phase space reconstruction techniques

A time series is a scalar sequence of observations Y_t that in itself does not properly represent the multidimensional space of the dynamical system, however we can think of it as a one-dimensional view of an unobserved process occurring in higher dimensions, from which the observed complexity or apparent randomness arises. Since our aim is investigating these underlying dynamics, getting information about the state space and the attractor that produced Y_t, we look at methods that allow reconstructing the state space from a single time series.

The most important technique, proposed by Takens [200], is the method of delays, or time-delay embedding: lagged coordinates are used to embed the observed time series in higher dimensions. Although embeddings could be created from as many time series as the dimensions of the state space, here we consider the worst case where only one time series is available. This reconstruction will not retrieve the original state space, but an approximation of it: something we shall call a *phase space*. Takens proved that a phase space obtained via time-delay embedding retains essential properties of the original state space (formally called topological characteristics), working as a surrogate for the state space of interest: it can be used to make forecasts, and information about the original attractor can be gained by exploring its properties. In the bottom panel of Figure 8.2 we can see a phase space of the Lorenz attractor reconstructed by embedding the time series $x(t)$ pictured in the middle panel.

To clarify embedding with an example, we take a time series $\mathbf{y_t} = \{y_1, \ldots, y_{10}\}$ and we set embedding dimension $m = 4$ and time delay $\tau = 2$: we make $m-1$ copies of the observed values $\mathbf{y_t}$ with a fixed time delay τ.

t	1	2	3	4	5	6	7	8	9	10
$\mathbf{y_t}$	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
$\mathbf{y_{t+\tau}}$	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}		
$\mathbf{y_{t+2\tau}}$	y_5	y_6	y_7	y_8	y_9	y_{10}				
$\mathbf{y_{t+(m-1)\tau}}$	y_7	y_8	y_9	y_{10}						

The m-dimensional reconstructed vectors that we obtain represents coordinates in the new 4-dimensional embedding space:
$\mathbf{r_1} = (y_1, y_3, y_5, y_7)$
$\mathbf{r_2} = (y_2, y_4, y_6, y_8)$
$\mathbf{r_3} = (y_3, y_5, y_7, y_9)$
$\mathbf{r_4} = (y_4, y_6, y_8, y_{10})$

But how do we choose m and τ? There are no mathematical rules for selecting the 'correct' values of the embedding parameters that ensure a faithful reconstruction of the phase space: the theoretical requirements are quite straightforward, however in practice estimating good values for these parameters from short and noisy time series can be challenging. Dozens of methods have been developed in the past few decades; we cover a few representative members of this set.

τ is usually chosen first. The original embedding theorems only require that τ is positive and not a multiple of any orbit's period, however this refers to the hypothetical situation where an infinite amount of noise-free data is available. In practice, when noisy, finite-length time-series data are available, a higher τ is needed to properly unfold the dynamics: too small delays would lead to strong correlation among the m coordinates in each of the reconstructed vectors, and so the embedded dynamics would lie close to the main diagonal of the reconstruction space. Since improperly unfolded reconstructions are not topologically conjugate to the true dynamics, this is a key problem [14].

While removing time-induced dependency is necessary, the smallest reasonable time delay τ must be chosen, as large τ could bring distant parts of the trajectory accidentally close together. A statistic that measures the independence of τ-separated points in the time series is generally computed: for example, the first zero of the autocorrelation function of the

time series would yield the smallest τ that maximises linear independence of the coordinates of the embedding vector. However, as autocorrelation only quantifies linear dependence, an alternative measure should be considered for nonlinear systems: mutual information is a measure of the mutual dependence between the two variables y_t and $y_{t+\tau}$, relating their probability distributions as follows:

$$I(Y_t, Y_{t+\tau}) = \sum_{y_t \in \mathfrak{Y}} p(y_t, y_{t+\tau}) log_2 \frac{p(y_t, y_{t+\tau})}{p(y_t)p(y_{t+\tau})} \tag{8.2}$$

After summarising $I(Y_t, Y_{t+\tau})$ for each choice of τ by taking the average mutual information (AMI) over the observed t, the optimal τ is selected as the first minimum of the AMI. When dealing with observed time series, rather than making any distributional assumptions, AMI can be computed using empirical probability distributions (estimated via density histograms).

After choosing a value for τ, the next step is to estimate the embedding dimension m. The original embedding theorems require $m > 2D$, where D is the unknown dimension of the underlying dynamics. In practice, as in the case of τ, we search for the smallest m that leads to a topologically correct result. Such a 'minimal sufficient embedding dimension' can be determined based on the false near neighbor (FNN) algorithm [110]: FNN are points that are close to each other in the phase space because they are simply close in time rather than because of the geometry of the phase space. The algorithm allows getting rid of them by progressively increasing the embedding dimension: the time series is first embedded with $m = k$ and each point's near neighbors are computed (spatial distance measured by Euclidean or maximum norm), then the embedding dimension is increased to $k+1$ and the near-neighbor calculation is repeated. If any of the neighbors in k dimensions are no longer neighbours in $k+1$ dimensions, that is taken as an indication that the dynamics were not properly unfolded for $m = k$. Noise also disturbs neighbor relationships, though, and thus can affect the operation of FNN-based algorithms [14].

8.2.3 Convergent Cross Mapping (CCM)

Although correlation is neither necessary nor sufficient to establish causation, it remains deeply ingrained in our thinking: one might be tempted to conclude that uncorrelated variables have no causal relation, as for the time series of the Lorenz system in Figure 8.1, whereas in fact nonlinear dynamics could be involved. Conversely, ephemeral or "mirage" correlations are common in even the simplest chaotic systems, but this does not imply causal-

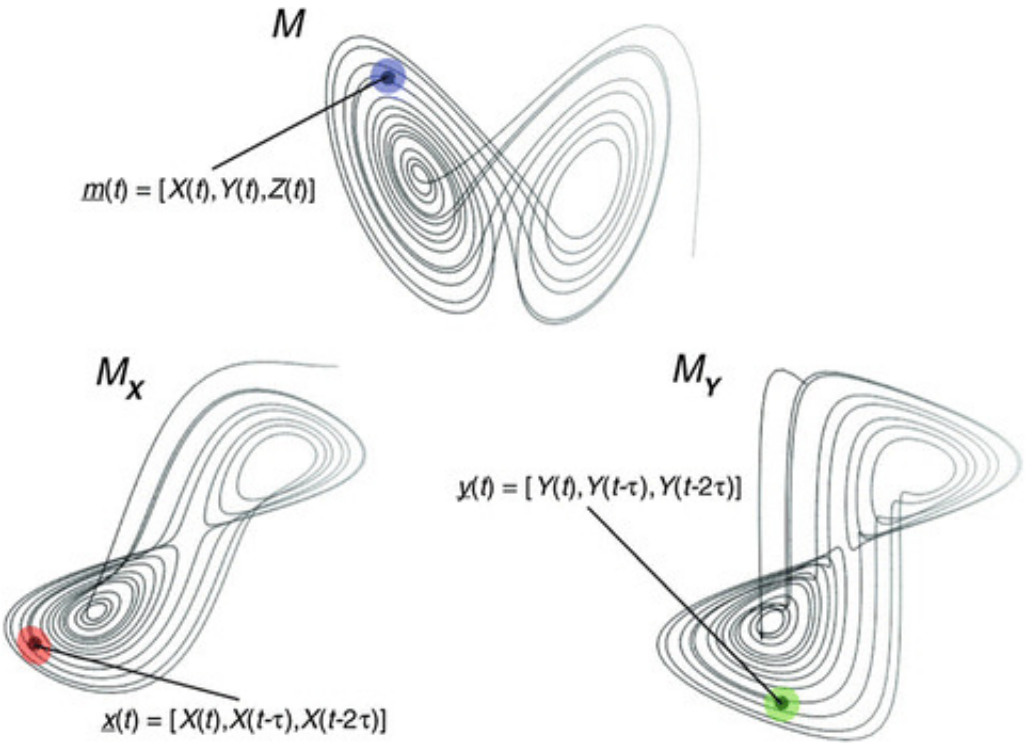


Fig. 8.3 Graphical representation of the correspondence of points on manifolds M, M_x and M_y as presented by Sugihara et al. [199]

ity. As discussed in section 3.4, Granger causality is a well-known data-driven method that allows inferring causality in stochastic systems, even if non linear: y Granger-causes x if the predictability of x decreases when y is removed from a stochastic model for x. However, if x and y deterministically interact in a dynamical system, the requirement of separability is not met as information about y is encoded into x, so Granger causality is not applicable.

In order to test whether two time series deterministically interact in a dynamical system, and are thus causally related, in 2012 Sugihara et al. [199] proposed CCM, a method based on time-delay embedding that exploits an extension of Takens' theorem. Takens showed that, given a dynamical system, a one-to-one mapping holds between its original manifold M and a shadow manifold M_x, an approximation of M obtained by embedding the $x(t)$ time series. However other manifolds reconstructed by embedding other time series originating from that system, for example M_y obtained by embedding $y(t)$, also map to M. Since both M_x and M_y map one-to-one to the original manifold M, they also map one-to-one to each other. This concept is summarised in Figure 8.3.

This idea was used by Sugihara et al. [199] to propose a *cross mapping* algorithm that aims to tests whether two time series $x(t)$ and $y(t)$ belong to the same dynamical system by testing the correspondence between phase spaces reconstructed from $x(t)$ and $y(t)$: if points that are nearby on M_y correspond to points that are nearby on M_x, then $x(t)$ and $y(t)$ originated from the same dynamical system. Detailed steps of this algorithm are enumerated below:

1. reconstruct manifolds M_x and M_y from time series $x(t)$ and $y(t)$ using time-delayed embedding
2. select a reference time t_k and identify the corresponding points on each attractor: $M_x(t_k)$, $M_y(t_k)$
3. compute Euclidean distances between $M_y(t_k)$ and all the points on M_y, and identify the $m+1$ nearest neighbouring points to $M_y(t_k)$, where m is the embedding dimension. Say $M_y(t_a)$, $M_y(t_b)$, $M_y(t_c)$ are selected if $m = 2$
4. compute relative distances of $M_y(t_a)$, $M_y(t_b)$ and $M_y(t_c)$ from $M_y(t_k)$ with respect to the total distances of all nearest neighbours, and call them e.g. d_{t_a}, d_{t_b} and d_{t_c}
5. use time indices t_a, t_b and t_c to identify the corresponding points on M_x, namely $M_x(t_a)$, $M_x(t_b)$, $M_x(t_c)$
6. cross map $\widehat{M_x(t_k)}$ as a locally weighted mean of $M_x(t_a)$, $M_x(t_b)$, $M_x(t_c)$ with weights d_{t_a}, d_{t_b} and d_{t_c}.
7. compare the estimated $\widehat{M_x(t_k)}$ and the actual value $M_x(t_k)$ by the Pearson correlation coefficient ρ in order to measure the cross mapping predictive skill

In summary, if nearby points on M_y help identifying nearby points on M_x, y can be used to cross map x, i.e. states of x can be estimated from records on y alone. Then, the ability to estimate the values of x from y can be seen as a measure of how much information about x has been encoded into y.

This procedure is iterated over the whole attractor by augmenting the learning set of one period at each step: with longer time series we expect reconstructed manifolds to be denser, nearest neighbours to be closer and cross-map estimates to increase in precision. Convergence, i.e. increase in cross-map precision as the portion of time series used to reconstruct M_y increases in size, was initially used as a practical criterion to detect causation.

This is why the method is called *convergent* cross mapping: if lengthening the training set for M_y led to increasing ability of y to cross map x, then this evidence would support the presence of an influence of x on y in the underlying dynamical system, i.e. a causal effect of x on y.

This theory implies that causal drivers will produce good reconstructions, but not that non-causal drivers will not produce good reconstructions. In particular, when the dynamics of a response variable y become dominated by those of the driving variable x, the full system consisting of x and y collapses to just that of x (system synchronized to the driver) and, although there is no causal effect of y on x, CCM is observed in both directions since the states of x can also uniquely determine y. This false positive result indicates that CCM may not be able to distinguish between bidirectional causality and strong unidirectional causality that leads to synchrony [199].

To circumvent this issue, a new criterion for causality has been proposed by Ye et al. [237]: since the effect of a driver x might be seen on a response y with some time delay, they extended CCM to include the directionality of information in time, i.e. y is used to cross map different lags of x. In the case of synchrony caused by strong unidirectional forcing of x on y, we expect y to be better at predicting the past values x (while x would best predict future ys), hence the true causal direction would be identified by an effective cross mapping for negative lags. A graphical explanation is presented in Figure 8.4. This method, that Cobey and Baskerville [28] renamed *negative cross-map lag criterion*, more generally improves the understanding of time delays in causation and helps identifying the correct ordering of variables in a transitive causal chain. Finally, it can also identify time-delayed interactions of stochastic drivers that have no dynamics [237]. CCM has been successfully applied in climate system, investigating the interaction between temperature and greenhouse gases [219], galactic cosmic rays and temperature variations [210], carbon cycle and tropical temperature variations[227] and soil moisture and precipitation [229].

In conclusion, informativeness of data is key to the success of such methods: if the time series is too short, cycles occurring at lower frequency might not be adequately sampled, and noisiness could obscure the resolution of the reconstructed system. To mitigate this problem, some pre-processing of the data is generally required: this allows separating observations into structured signal and unstructured noise. If a strong signal is detected, then EDM methods can be applied on it; on the other hand, if signal is weak, conventional linear stochastic models may be more effective [97].

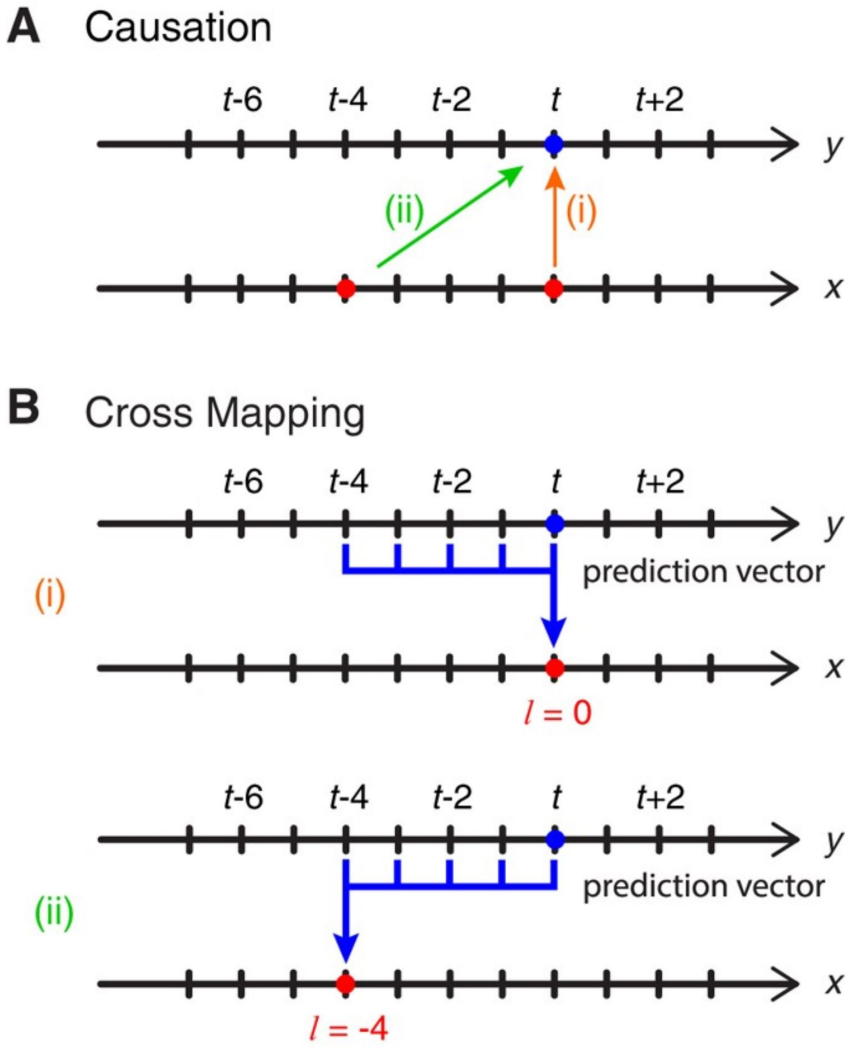


Fig. 8.4 Effect of time delays on cross mapping. Panel (A) shows causation for two cases: (i) no time delay in the effect of x on y and (ii) y responds to x with a time delay of 4 time steps. Panel (B) shows (i) cross mapping with l=0, equivalent to the original formulation by [198] and (ii) cross mapping with l=-4, which may be expected to be better than l=0 when x acts on y with some time delay. [237]

8.2.4 Signal processing techniques

As anticipated in section 8.1.1, when approaching NLTS analysis we primarily want to assess whether the data, including its observed irregularities, should be attributed to noisy (linear) behaviour or to deterministic (nonlinear) dynamics. Such noise can be defined as unstructural variation that doesn't evolve systematically with time, and its presence can be an important obstacle to identification of deterministic dynamics [214]: EDMs presented in section 8.2.2 can successfully detect the shape of nonlinear systems once noise has been eliminated.

Thus, data pre-processing techniques play a key role in unmasking components that are more likely to describe the underlying system: they typically decompose the original series into a sum of a small number of independent and interpretable components, such as a slowly varying trend, oscillatory components and structureless noise, that will keep invariant characteristics over time, i.e. $X(t) = X_1(t) + X_2(t) + \cdots + X_r(t)$. Success of this process relies on the property of separability of these components, i.e. the possibility of extracting them from the observed sum $X(t)$ [72].

Among many methods we present **singular spectrum analysis (SSA)**, a data-driven signal processing technique able to detect structural variation in irregular time series data without imposing theoretical assumptions on the source of irregularity [72]. For this reason, it is also referred to as a model-free approach. Although not very popular among statisticians, SSA is widely used in physics, meteorology and climatology: it incorporates elements of classical time series analysis, multivariate statistics, multivariate geometry and linear algebra. We choose SSA because it is particularly suitable to separate trend and seasonality, e.g. by identifying underlying sine waves with different frequencies. Moreover, Hassani [79] compared its performance to alternative methods such as Box-Jenkins SARIMA models, the ARAR algorithm and the Holt-Winter algorithm, and concluded that SSA is more accurate than several well-known methods in terms of forecasting results.

The algorithm of SSA consists of two complementary stages: decomposition and reconstruction. Consider a time series $x_t = (x_1, \ldots, x_N)$, and let L ($1 < L < N$) be some integer called the window length. Decomposition starts with embedding the original time series x_t with delay $\tau = 1$ to obtain $K = N - L + 1$ lagged vectors of size L that, combined by column, form a trajectory matrix $\mathbf{B}$ of size L x K. Such a trajectory matrix is then subjected to singular value decomposition, explained in detail in appendix C.1. In summary, decomposition of matrix $\mathbf{B}$, of rank r, results in the sum of r matrices of rank 1, called elementary matrices, ordered from the largest to the smallest. Each of these matrices can be identified by a vector

of length 3, called an eigentriple, comprising the i^{th} singular value, the i^{th} eigenvector of $\mathbf{BB}^T$ and the i^{th} eigenvector of $\mathbf{B}^T\mathbf{B}$.

Reconstruction follows: firstly, the matrices obtained in the decomposition stage are assigned to several disjoint groups, a step called eigentriple grouping, and summed within group. Finally, in the last step each resultant matrix is transformed, via diagonal averaging, into a time series.

Some recommendations on choosing values for the parameters that maximise separability will follow. Firstly, better separability is obtained when L is large enough ($L \sim N/2$) and, if we want to extract a periodic component with known period (e.g. 52 weeks), then L should be divisible by that period [72]. Secondly, choosing an optimal grouping for eigentriples is crucial for a proper decomposition of the observed time series: the aim is distinguishing eigenvectors that represent signal from the ones that represent noise, and grouping eigenvectors that describe similar behaviours. For example, any two sine waves with the same frequency and phase shift, generated by the same sinusoid, should be grouped to represent a given periodicity of seasonality.

Several leading eigentriples should be selected and visually inspected. The spectrum of the singular values, i.e. the plot of the singular values against the index of the eigentriples, is a good starting point to discriminate the nature of each component: singular values of eigentriples of harmonic series are usually very close to each other (paired), hence they can be easily spotted by a plateaux in the plot, whereas a slowly decreasing sequence of singular values generally indicates noise Further, each eigenvector can be converted into an elementary reconstructed time series, and both its visualisation and a matrix of weighted correlations between them (w-correlation matrix) are helpful for separation: we want to make sure that correlated components are not included into different groups, hence we should focus on small correlation, indicating well separated components (if correlation is zero they are said to be w-orthogonal).

Finally, graphs of eigenvectors can also help in the process of grouping, as their form replicates the form of the time series component that produces them: for example, one or more of the leading eigenvectors might be slowly varying, with no oscillatory behaviour, hence representing the time series trend. The scatterplot of pairs of eigenvectors, which produces a more or less regular T-vertex polygon, can help to identify a sinusoid of period T [71].

8.3 Application to influenza and IPD time series

8.3.1 SSA for IPD

We consider the time series presented in chapter 4.2, Figure 7.8, i.e. the weekly IPD incidence rate per 100,000 residents in England from 2009 to 2018. We start our analysis by running a singular spectrum analysis to separate unstructural noise from signal. For better separability, we select $L = 208$ as the maximal window length such that $L \leq$ N/2 and L is divisible by 52. We obtain results using the `Rssa` R package.

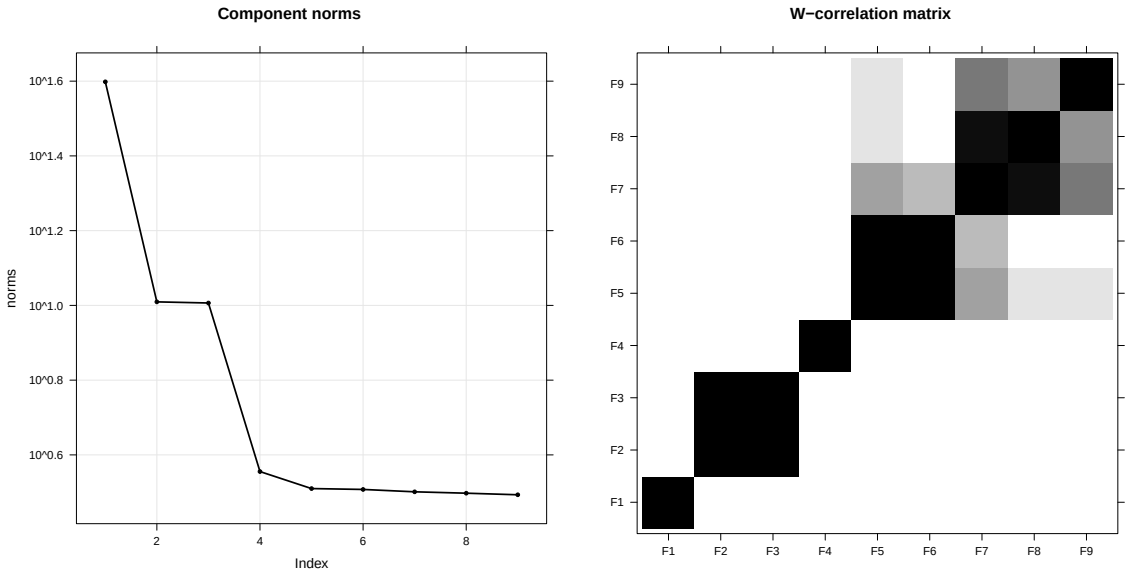


Fig. 8.5 Eigenvalue spectrum and W-correlation matrix for IPD

The ten leading eigentriples are selected and inspected. The spectrum of singular values is presented in the left panel of Figure 8.5: one evident pair corresponds to eigenvectors 2 and 3 of the series, suggesting an important role played by a sine wave. Two important jumps divide eigenvectors 1 and 4, probably trend components, whereas the slowly decreasing singular values for component 5 and upwards suggest the presence of noise. A good separability for components 1 to 4, and a bad separability for components 5 to 9, are shown by the w-correlation matrix in the right panel of Figure 8.5, where the grey scale from white to black corresponds to values of the correlation between eigenvectors from 0 to 1. Series 1 to 4 are not correlated with any other one (w-orthogonal) and components 2 and 3 are considered as one group.

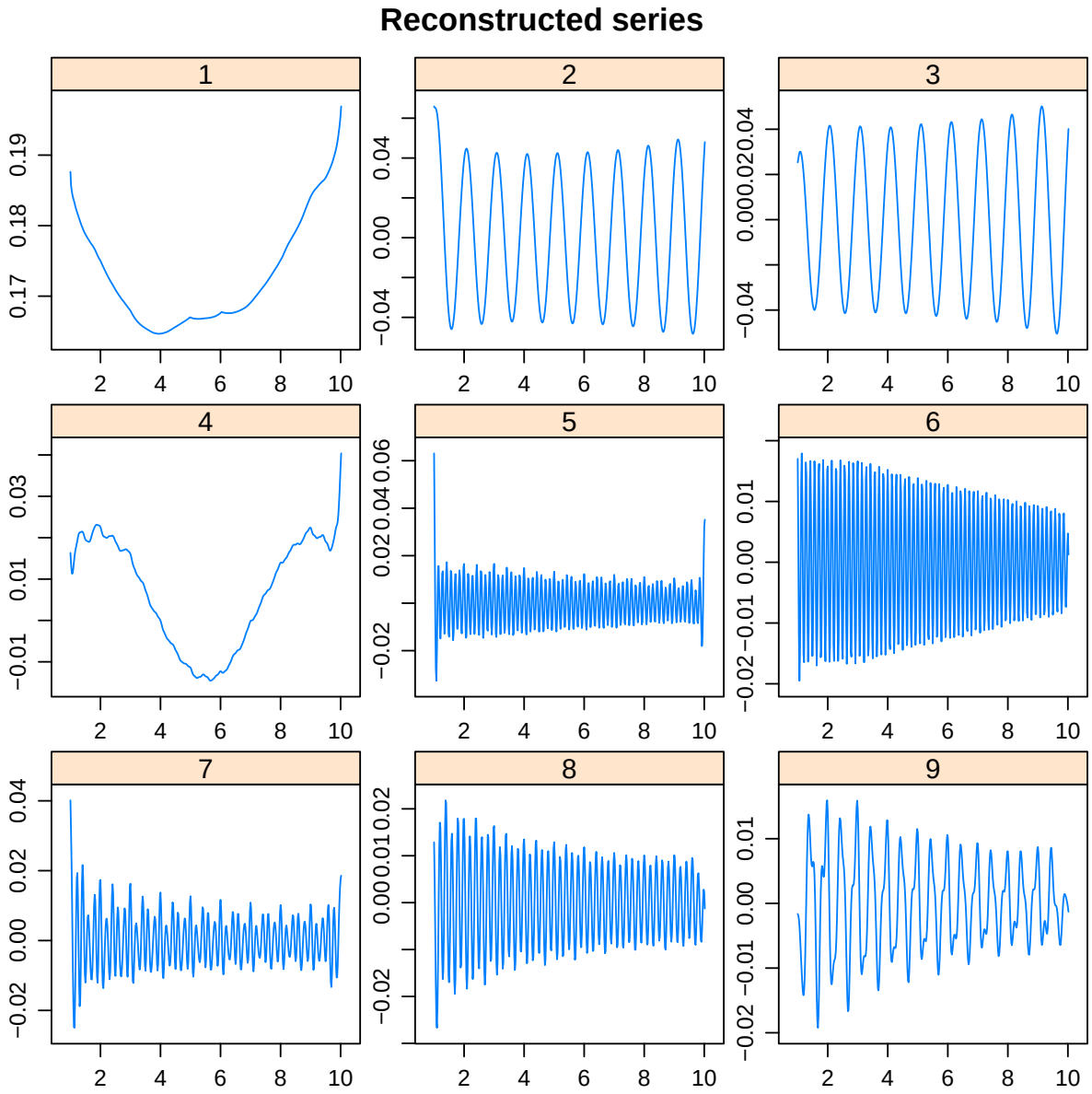


Fig. 8.6 Reconstructed elementary components for IPD

The speculated forms of leading eigenvectors 1 to 4 are confirmed in Figure 8.6, which represent reconstructed time series from each elementary component: 2 and 3 are high-frequency components which oscillate at the same frequency, representing yearly seasonality, while 1 and 4 are trend components. Single and paired eigenvectors plots are listed and described in C.2, Figure C.1.

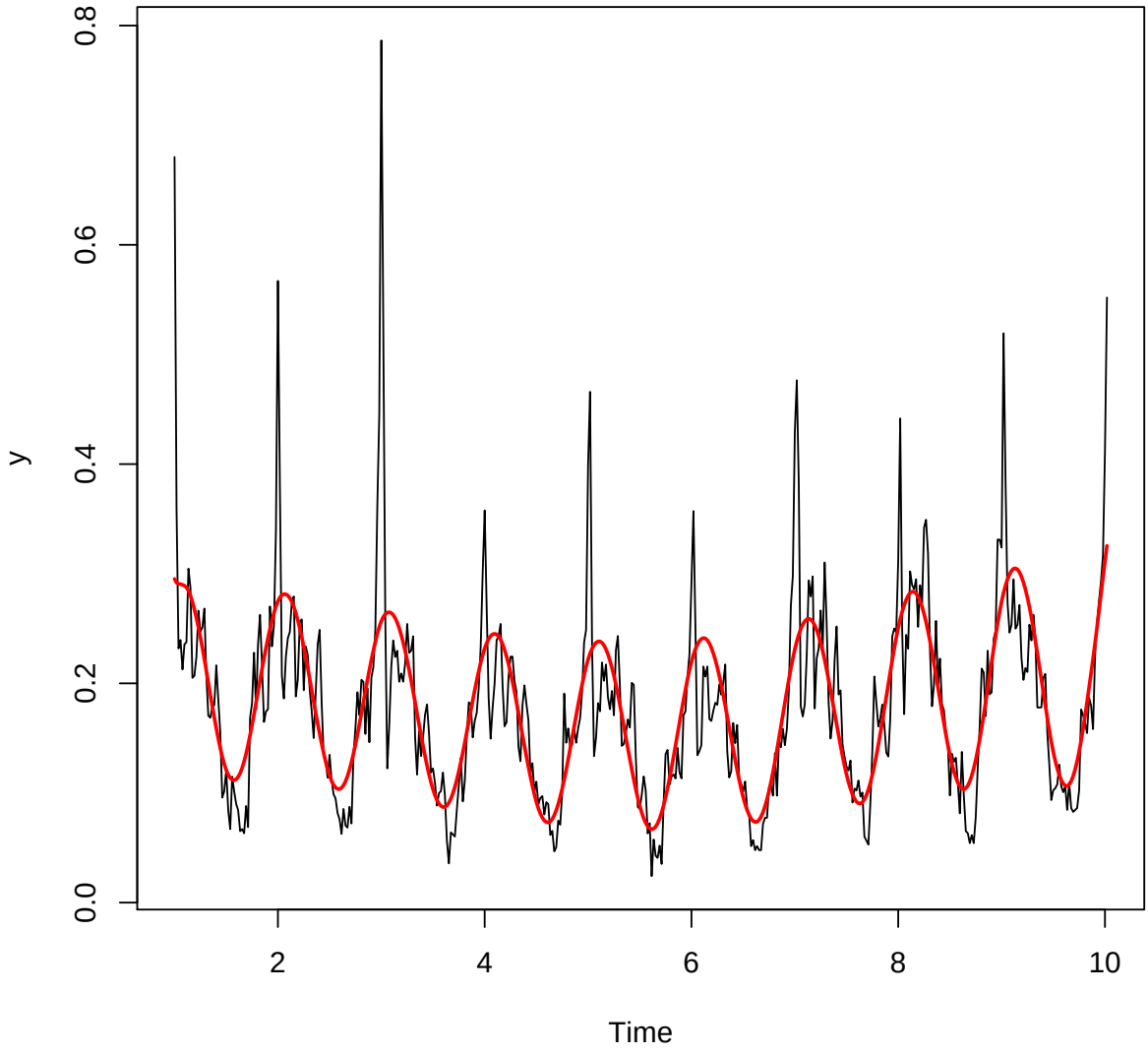


Fig. 8.7 Signal for IPD reconstructed with SSA

Hence, we extract IPD signal as the sum of components 1 to 4, as shown in Figure 8.7: the extracted trend+seasonality is superimposed in red on the background of the original series, and it explains 90.3% of the variability. We also run a second SSA on the residuals to check if other eigenvectors could be extracted, but the w-correlation matrix confirms no further separability (Figure C.2).

8.3.2 SSA for influenza

As for IPD, we consider the influenza time series presented in chapter 4.2, the weekly incidence rate per 100,000 residents in England, and we run SSA with $L = 208$. The spectrum of the ten leading singular values is presented in the left panel of Figure 8.8: as in the case of IPD, two clear jumps divide eigenvectors 1 and 4, probably trend components, singular values for components 2 and 3 are very similar, indicating harmonic series, while singular values slowly decrease from eigenvalue 5 upwards, hinting at noise. From the right panel of Figure 8.8 we see how separability is perfect for components 1, and 2-3 as a group, whereas the w-correlation matrix has a lot of grey for components 4 upwards.

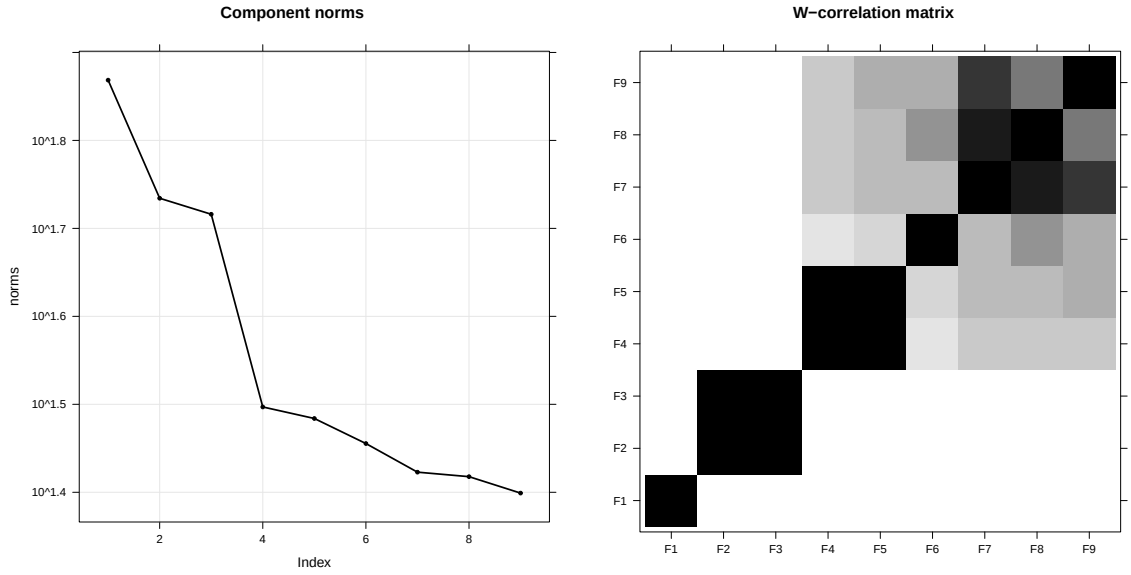


Fig. 8.8 Eigenvalue spectrum and W-correlation matrix for influenza

Time series reconstructed from each elementary component, in Figure 8.9, confirm an upwards trend in component 1, oscillatory behaviour of components 2 and 3, with the same frequency but phases slightly shifted, and noisy dynamics for components 4 to 9. Single and paired eigenvectors plots are listed and described in C.2, Figure C.3.

We conclude SSA by extracting influenza signal as the sum of components 1 to 3, as plotted in red on the background of the original series in Figure 8.10: this time the selected components only contribute 42.6% of the variability, leaving a lot of the dynamics attributed to noise. However, a second SSA confirms no further separability in the residuals (Figure

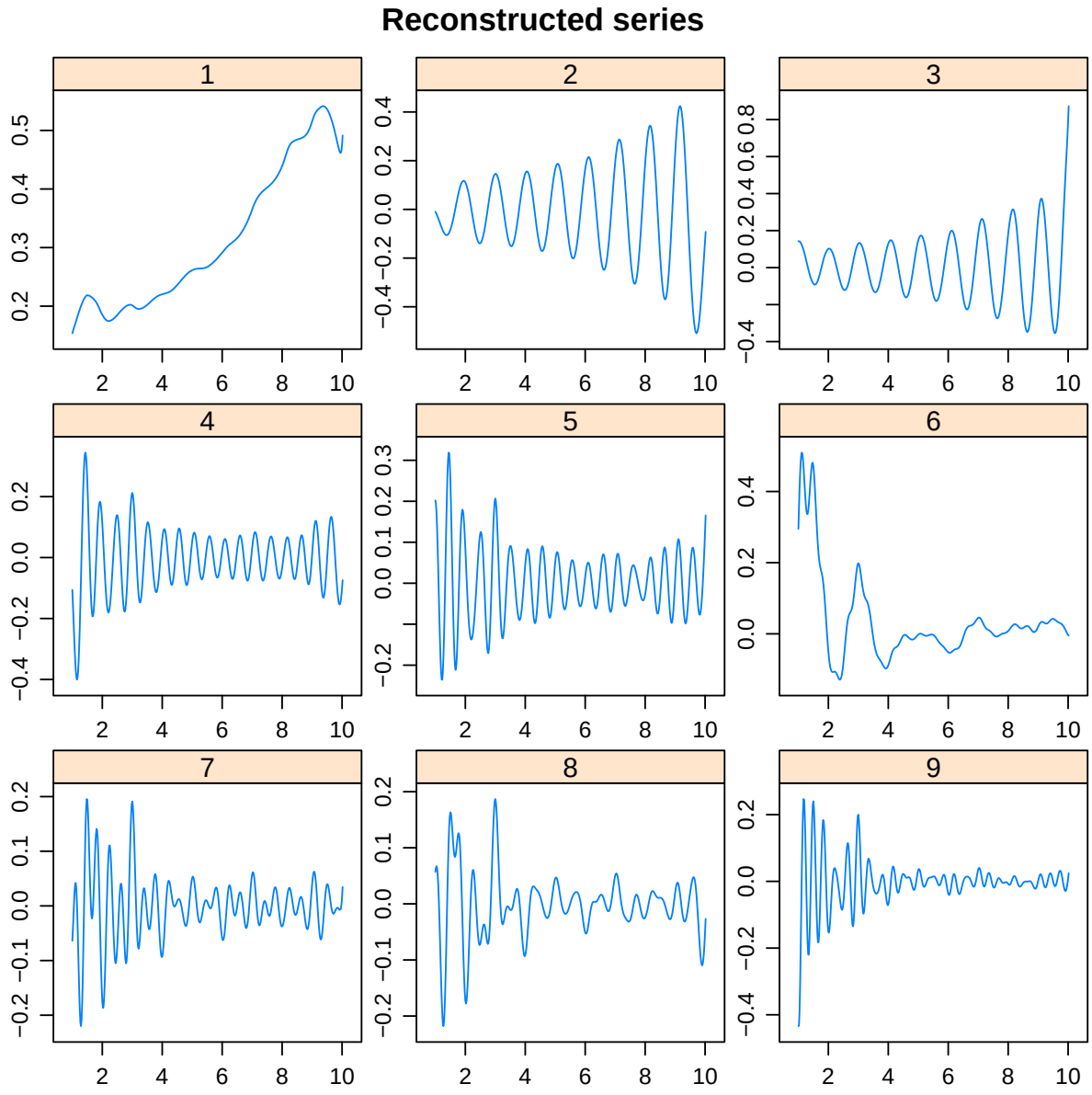


Fig. 8.9 Reconstructed elementary components for flu

C.4).

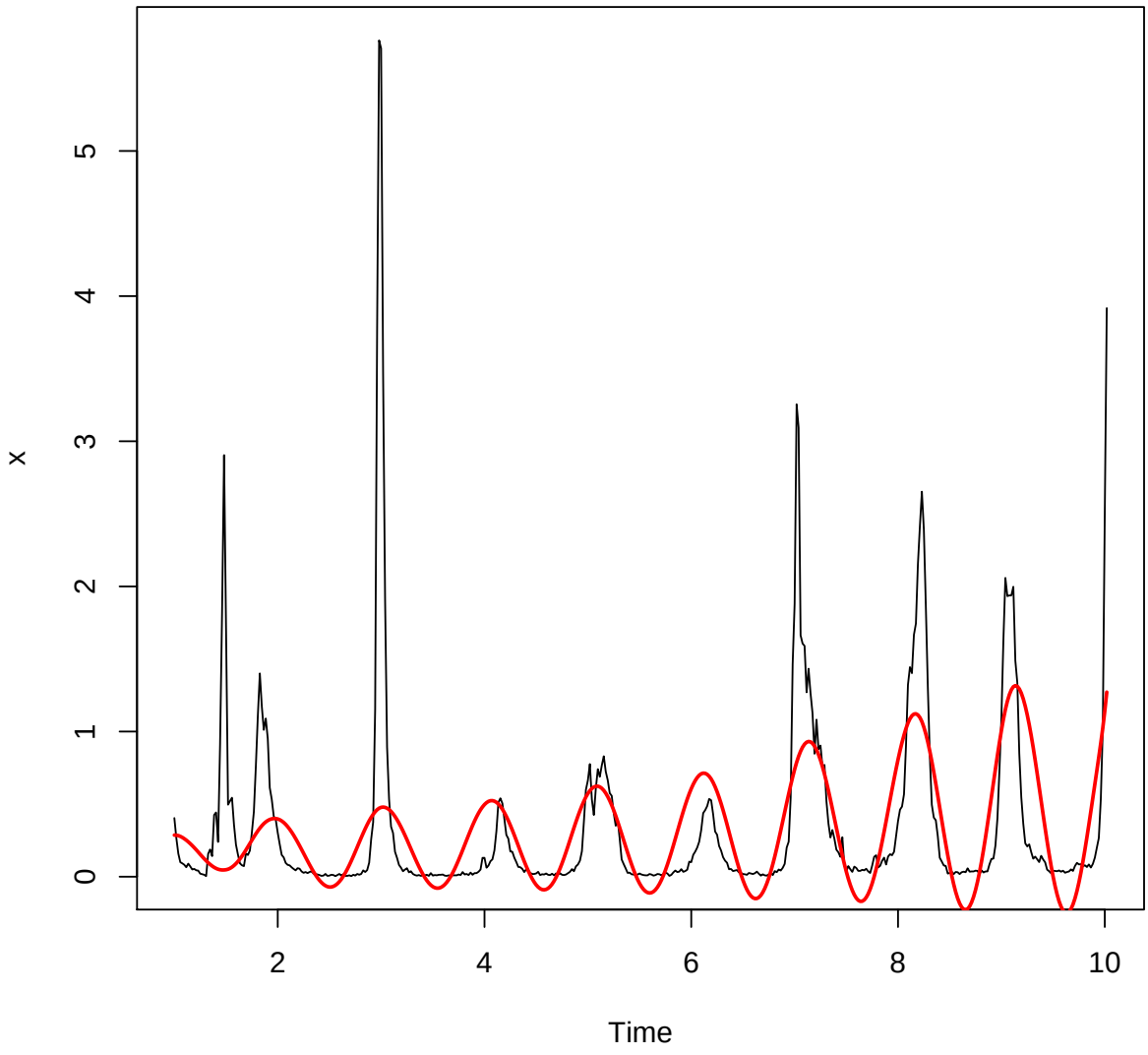


Fig. 8.10 Signal for influenza reconstructed with SSA

8.3.3 Embedding of IPD and influenza signals

In order to reconstruct the underlying dynamics of IPD and influenza, we first identify the best embedding parameters for the respective extracted signals. The time delay, τ, is selected based on minimum mutual information, whereas the embedding dimension E is selected as the first minimum in the percentage of false nearest neighbours when increasing embedding dimension from $m-1$ to m, i.e. when the proportion of points that were identified

as neighbors in dimension $m-1$ and are no longer neighbours in dimension m is minimum. The chosen parameters to embed IPD are τ=16 and E=3, as shown in Figure 8.11, whereas for influenza we find τ=11 and E=3 (Figure 8.12).

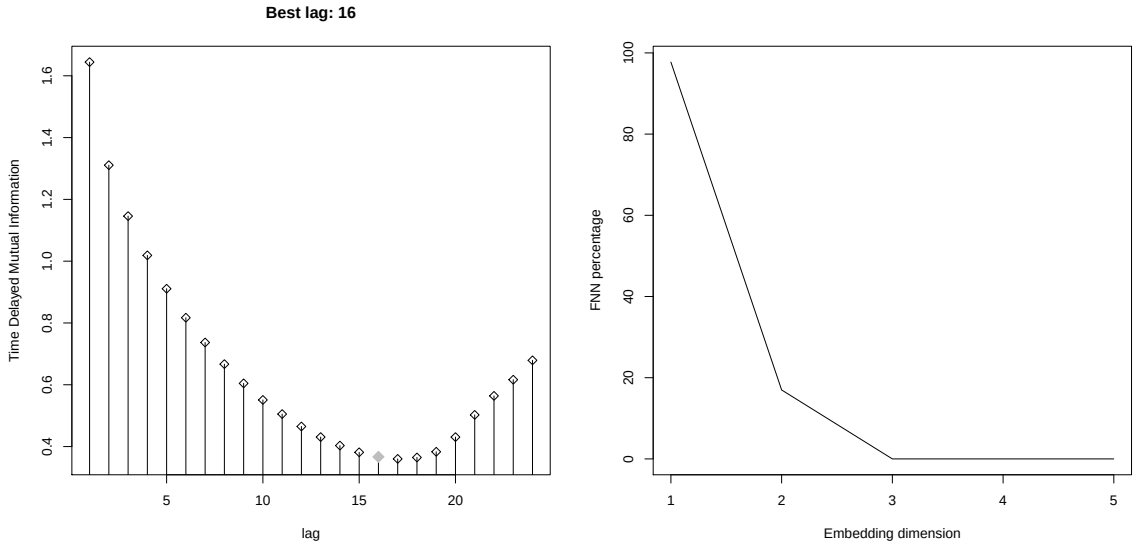


Fig. 8.11 Selection of time delay τ and embedding dimension E for the IPD signal

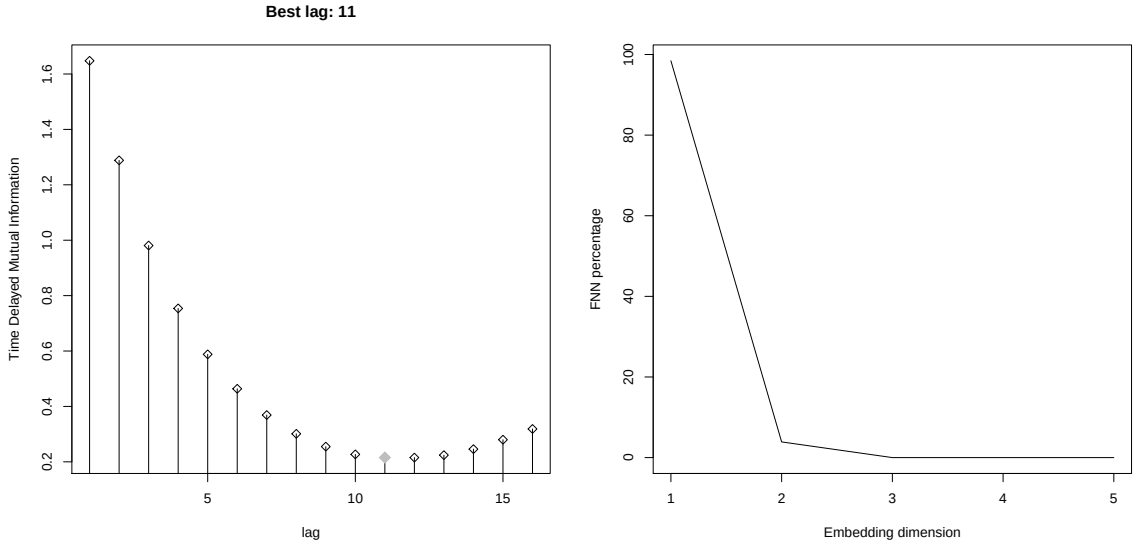


Fig. 8.12 Selection of time delay τ and embedding dimension E for the flu signal

For comparison, in section C.2.3 we present the embedding parameters that would have been chosen if we had computed AMI and FNN for the original rates instead of the extracted

signals: in both cases, we would have required a higher embedding dimension: 4 in the case of IPD, 5 in the case of flu. Finally, 3d scatterplots of the corresponding manifolds are presented in Figure 8.13: both of them show a spiral trend, however the time series are too short to clearly identify the underlying attractor.

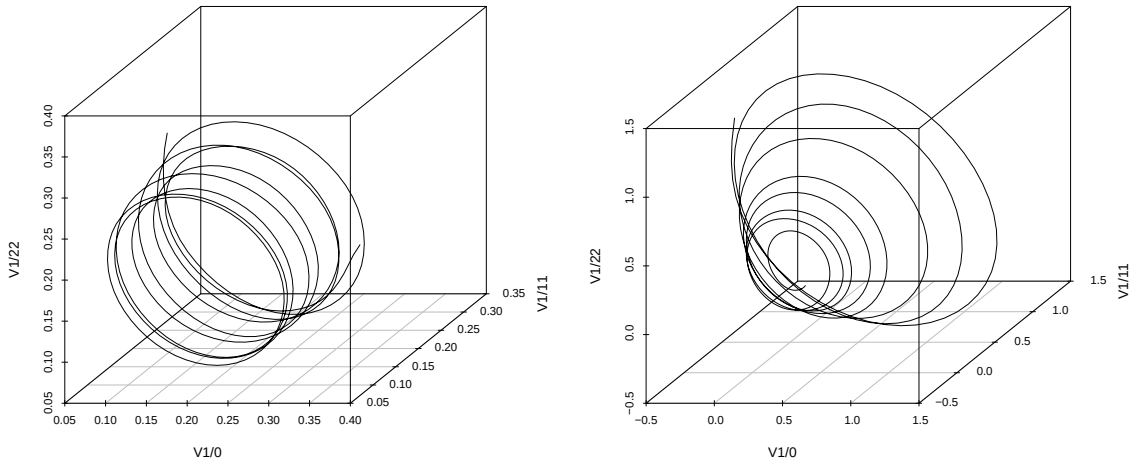


Fig. 8.13 Reconstructed manifolds for IPD signal (left) and influenza signal (right)

8.3.4 CCM of influenza and IPD

Despite reconstructing two separate state spaces, our question concerned whether the two phenomena did belong to the same dynamical system. CCM allows testing for the presence of a causal relation between influenza and IPD by using nearby points on one attractor, e.g. M_{IPD}, to map nearby points on the other, e.g. M_{flu}. In order to demystify the ordering of variables in the causal chain, IPD can be used to predict different lags of flu as, if flu is a driver for IPD, its effect on IPD might be seen with some time delay, and vice versa. Hence, we cross-map IPD from influenza, and influenza from IPD, considering lags from -5 to +5 weeks.

The cross mapping predictive skills, expressed as correlation ρ between the estimated and observed values, are summarised in Figure 8.14: we find that correlation of flu xmap IPD is larger (ρ above 0.9 for any lag), meaning that states of IPD can be estimated very well from records on flu alone; however, the converse is also true, as the correlation for IPD xmap flu is also above $\rho = 0.75$ for any lag. Since the ability to estimate values of one variable

from the other can be seen as a measure of how much information is encoded in that variable, we can conclude that both variables encode information about the other one.

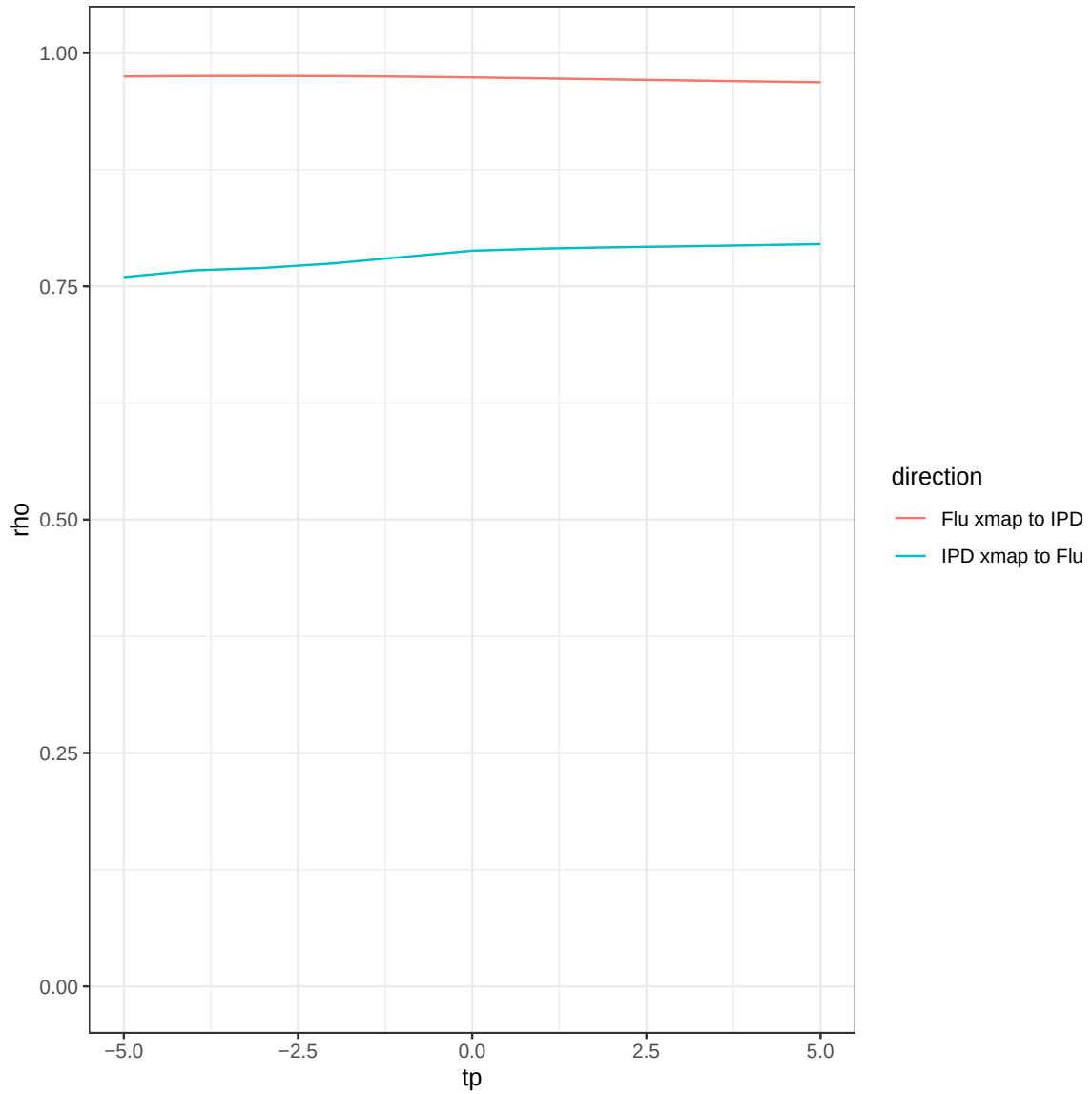


Fig. 8.14 Correlation coefficients for prediction skills of time-delayed CCM

Directionality of information in time also does not help greatly: flu better cross-maps negative lags of IPD, suggesting a potential driving role, but the difference in predictive skill across lags is really small, so we do not feel this provides evidence towards a strong unidirectional forcing of IPD on flu.

For comparison, we repeat the same analysis on the original time series, and summarise cross-map predictive skills in Figure C.7: cross-mapping skills are significantly lower when including the component identified as noise in SSA, yet a clear maximum in correspondence of no lag is visible, suggesting no true causal relationship between influenza and IPD but rather a synchrony due to some shared driving factors.

8.4 Discussion

The collection of ideas and techniques known as EDM can be extremely effective when the data derive from deterministic dynamics in some state space. This analytical framework allows to solve an inverse problem of considerable complexity: from data you can infer properties of some hidden dynamical system. However, the assumption of determinism makes these methods largely unsuitable for characterizing stochastic aspects of data.

All the methods considered here rely on reconstructing the unobserved state space thanks to the powerful method of delay-coordinate embedding. Yet, this method has some practical limitations. Firstly, estimation of the minimal sufficient embedding delay and dimension is a non-trivial problem [17], as thoroughly described in section 8.2.2. Further, resolution of such an attractor might be obscured when short and noisy time series are available. Data pre-processing techniques have been proposed to extract the high-frequency components of the signal, yet distinguishing noise from deterministic chaos can be a harder challenge as both types of signals exhibit irregular temporal fluctuations, and there is a risk that these algorithms might filter signal out along with the noise [202].

We applied SSA to extract signal from observed time series data of influenza and IPD incidence in England, and used the resulting signal-processed time series to reconstruct the underlying systems. Trend and seasonality components were isolated and from both time series: these successfully explained over 90% of the variability for IPD, whereas 57% of influenza variability was left unattributed. We might argue that, despite identifying trend and seasonality, our algorithm excessively smoothed the observed time series: by looking closely at the unattributed components we see how some important information in the data, such as the different phase of seasonal spikes, has been filtered out. Moreover, embedding of these extracted signals resulted in well-defined spiral-shaped manifolds for both phenomena, as expected due to annual seasonal oscillations, however the limited length of our time series prevented us from detecting a real convergence to an underlying attrac-

tor, leading to inconclusive evidence about the deterministic nature of the considered systems.

Further, two complex systems that appear to be weakly coupled are difficult to analyze, as there are multiple ways in which they could be causally linked: neither influences the others' temporal dynamics, and the variables are therefore causally unrelated; a forcing process influences the temporal dynamics of a response process, but the response process has no effect on the forcing process in return, i.e. there is unidirectional causality; or there is bidirectional causality where each variable influences the others' dynamics. Standing on the theory of embedding, CCM has recently been introduced as a practical numerical approach to identify causal relationships in weakly coupled nonlinear systems without requiring correlation. However, its ability to distinguish among these three cases depends on the strength of underlying relationship, beyond the general issues with state space reconstruction described above [26].

We applied time-delay CCM both to the signal-processed and the original time series to test for the presence of any causal relationship between influenza and IPD: each variable contained information about the others' dynamics, suggesting they could interact in the same dynamical system, however lack of a clear temporal ordering did not allow us to conclude towards any directionality. We replicated this analysis both on signal-processed and noisy data as our aim was not to distinguish between chaos and noise, but rather attempting to reduce the impact of noise when gaining insights on the underlying system complexity.

While we acknowledge that pathogens of interest might interact in a nonlinear way, the complexity and multidimensionality of the setting, including environmental forces that have not been inspected here, does not lead us to recommend an EDM analysis to make statements about their causal interaction: as Cobey and Baskerville [28] highlighted in their work, the little prior knowledge of a system's complexity, with transient dynamics and noise, prevents us from reaching statistically rigorous conclusions about who interacts with whom. Further, this method does not allow the inclusion of covariates, making it impossible to account for meteorological effects as we did in previous chapters. Finally, incidence for the pathogens of interest could be effectively described by a nonlinear system to produce satisfactory short-term forecasts, but identification of long-term properties remains an open question. Finally, the challenge of separability and noise isolation also requires more work.

Chapter 9

Conclusions and future work

Over 100 years have elapsed since the 1918 influenza pandemic, yet the impact of respiratory pathogens on human health remains substantial despite increasing availability of novel preventative and therapeutic strategies. Efforts to enhance pandemic preparedness should not be forgotten as the COVID-19 pandemic is vivid in our minds: novel pathogens continue to emerge in animals, subsequent spillovers into human populations are likely, and current global conditions, characterised by international travel, migration, and urbanization, enable pathogens to spread widely and quickly [126].

Even if a large portion of the LRTI deaths due to endemic pathogens could be avoided by enhancing access to healthcare for people who currently don't receive immunisation or timely antibiotics [150], the increasing incidence of chronic illnesses in an aging population and the rising number of drug-resistant pathogens place individuals at greater risk of infection and complications from respiratory viruses [24]. Further, declining levels of protection from vaccines due to anti-vaccination sentiments in some communities is enabling previously declining respiratory viruses to cause significant outbreaks [WHO].

Surveillance data represent a fundamental component in the monitoring of epidemics, both to inform detection of new outbreaks and retrospective reconstruction of past ones, and statistical methods play a crucial role in quantifying the burden and estimating effectiveness of interventions. The work of this thesis, placed at the interface of public health, epidemiology and statistics, is an example of what role these quantitative methods can play in fostering this understanding.

Regression-type approaches have been historically employed at Public Health England to provide a picture of disease incidence from surveillance data. We consider these methods to

be somehow too simplistic, and propose a portfolio of novel time-series-based approaches which allow modelling surveillance data while accounting for temporal dependency, approximating the transmission dynamics. Special attention is also being paid to an accurate quantification of uncertainty. Application of these new methods has the potential to improve substantive evidence, to inform policy and the government decision-making.

9.1 Main thesis findings

9.1.1 Pneumococcal disease progression

Chapter 2 makes use of a longitudinal study jointly tracking asymptomatic carriage of S.Pneumoniae and LRTI infections at the individual level. We reconstruct within-subject disease dynamics and quantify the impact of seasonality, temperature and viral circulation on the probability of disease progression and clearance. Importantly, our model estimates that the risk of developing pneumonia is 67% higher during the influenza season, and 47% higher during the RSV season, after taking into account meteorological variations.

The nature of this information, where the asymptomatic condition is ascertained at predefined time instances while time of LRTI occurrence is exactly identified as people seek healthcare in the presence of symptoms, posed some challenges to our inference. In particular, we used standard methods to account for exact observation time for the pneumonias, however we could not use the viral positivity measured in the cohort as a covariate, as it was only recorded at the pneumonia onset, being an informative observation. Hence, we replaced this with viral positivity from the national-level surveillance. A similar modelling strategy could be applied to other cohort studies tracking disease evolution of pathogens characterised by an asymptomatic phase.

9.1.2 Burden of endemic respiratory pathogens

Chapter 3 reviews currently used methods for the estimation of seasonal disease burden. The focus of the chapter is on the limitations of these models, highlighting in particular the need to account for time dependence and the challenge of disentangling the contribution of different factors with overlapping seasonality. Methods that include virological circulation are considered, as they represent a natural framework to estimate the counterfactual, i.e. the

baseline burden not attributable to a specific pathogen.

In chapter 4 we apply one of these methods to English surveillance data to clarify the contribution of the influenza virus on severe pneumococcal infections across age groups, in both seasonal and pandemic settings. We find their association to be significant, particularly in younger age groups, during the 2009 flu pandemic, whereas the seasonal contribution does not appear to be relevant. These findings have implications for pandemic preparedness in terms of advising on antibiotic stockpiles in England, for which currently there is no clear evidence.

We believe our approach could be valuably applied to retrospectively investigate relationships of other notifiable diseases. For example, the contribution of viruses to secondary bacterial infections due to *Staphylococcus Aureus* and *Streptococcus Pyogenes* requires further investigation, to better inform antibiotic prescription policies. Despite limitations due to the available data and modelling assumptions, the proposed model successfully improves existing understanding of interaction between multiple pathogens. A similar modelling strategy could be usefully employed by many countries that rely on infectious disease surveillance for informing policy, and extended to tackle spatial dynamics if region-specific counts are available.

9.1.3 Evaluation of pneumococcal vaccine and serotype replacement

Chapter 5 introduces the intervention evaluation framework. Estimation of the impact of an intervention is a key problem in many public health institutions and international organisations. We present two classes of methods, case-only and case-controlled, that have been used to estimate intervention effects from time series data. We focus initially on the ITS framework: this method relies on a pre-chosen function for the intervention effect, and suffers from unobserved confounding, however it is the best choice when no controls are available.

We then present the CIM method, a DLM including a Bayesian variable selection strategy, which allows including several control time series as covariates naturally dealing with multicollinearity issues. In fact, the spike-and-slab prior selectively includes covariates which maximise model fit to observed data, and the posterior distributions of the coefficients naturally yield weights to combine controls into a weighted average. Further, the DLM structure, describing the outcome dependency over time through an evolution equation, effectively accounts for temporal dependence avoiding the common assumption of independence

between observations.

We explore their application in chapter 6, investigating the impact of pneumococcal vaccine introduction in England. PHE have previously used counterfactual analysis somewhat parsimoniously in evaluating policy impact, relying on simple impact indicators [113]. Our work adds to such existing evidence: we successfully disentangle vaccine effectiveness from serotype replacement across age groups by modelling serotype-specific IPD incidence over an 18 year period, providing important evidence to inform future vaccination policies.

9.1.4 COVID-19 excess mortality

We cast the problem of estimating baseline mortality during an outbreak in terms of counterfactual estimation, following the methods presented in chapter 5, and we apply this idea to estimate excess mortality in England during the COVID-10 pandemic. Baseline mortality in the absence of COVID-19 is forecast from a DLM model fitted to the daily mortality rates observed for the 2019/20 epidemic year, before the pandemic started. Several predictors of all-cause mortality are considered, including mortality in the corresponding period of past seasons, as well as meteorological conditions and viral circulation. As described above, the most suitable predictors are selected through a spike-and-slab prior.

We estimate a cumulative excess of 100.8 (95% CrI 95.8-106.4) deaths per 100,000 residents from the 2nd March until 29th May 2020, 147% (95% CrI 145%-150%) of the mortality expected during a corresponding spring period. Important differences are identified across regions, with the excess mortality above the baseline estimated to be 178% (95% CrI 175%-182%) in London, and much lower than average in the South West, East Midlands and East of England.

Beyond quantifying excess mortality across regions and age groups, we identify differential starting dates for the outbreak across population subgroups making use of the Bayesian credible intervals for the estimated cumulative excess. Important differences can be seen across age groups in terms of when the excess deaths started: as early as 14th March in the age group 25-44, and as late as the 27th March in the 75-84 years old.

9.1.5 Empirical dynamical modelling

Finally, in chapter 8, we explore empirical dynamical modelling, which aims to investigate if one or multiple observed time series that have complex, highly variable appearance have been generated by an unobserved dynamical system, and to test whether two variables belong to the same dynamic system. This relates to chaos theory, which tries to reconstruct long-term dynamic behaviour of a system, exploring its potential nonlinear shape and whether it will settle to a steady state.

We applied time-delay CCM, a phase-space reconstruction method, to test for the presence of any causal relationship between influenza and IPD. We concluded that each variable contained information about the others' dynamics, suggesting they could interact in the same dynamical system, however lack of a clear temporal ordering did not allow us to make definitive conclusions about the direction of their relationship.

9.2 Future work

This thesis, while providing some answers to the questions posed at the beginning of the PhD, opens new research directions. Some of the future work should be aimed at extending and improving the models proposed in this thesis, as well as focussing on new methodological avenues.

9.2.1 Extensions to the hhh modelling framework

Nonlinearity of effects

In the *hhh* modelling work we assume effects of covariates to be linear, including possible lags. We then go as far as exploring nonlinear dynamics through empirical dynamical modelling and state-space reconstruction. Alternative solutions to investigate nonlinearity of effects could be considered among stochastic models. In particular, Gasparrini [61] proposed distributed lag non-linear models (DLNM), which allow nonlinear exposure–response dependencies and delayed effects, assuming the outcome to follow a Poisson distribution with independence among observations. If instead an autoregressive component should be included, the (G)ARCH-family of models, popular in the financial literature, could be

considered [52].

Time varying parameters

In the hhh model for IPD counts, we estimated the autoregressive coefficients to be constant over time. Such an assumption has kept our model easy to interpret and avoided overfitting, however this implies that both pneumococcal transmission and its interaction with influenza have no seasonal behaviour. This might not be true, as *S.Pneumoniae* transmission is likely to change over time following known and unknown factors, including climatic influence and host susceptibility. An extension of this work could account for difference in transmission across seasons, during school opening etc, making the proposed model more realistic, and providing a better approximation of pneumococcal transmission.

9.2.2 Incorporating a transmission model

Microbiologically confirmed cases of *S.Pneumoniae*, recorded in the surveillance system, do not fully capture the burden of pneumococcal disease, as only invasive disease confirmed by culture from sterile sites is identified. We use this information, assuming pneumococcal infection to follow from transmission, yet we are aware that pneumococcus can be carried asymptomatically, and that those individuals also contribute to the transmission.

Here we limited our work to phenomenological models, however an extension could involve mechanistic models, that allow modelling transmission and severity, along the lines of Opatowski et al. [162], Trotter et al. [207], van Hoek et al. [217]. Any information about carriage would prove useful in estimating transmission parameters. Further, if serotype-specific information was available both at the carriage and at the invasive disease stages, this would allow disentangling differential transmission and invasiveness of each serotype.

9.2.3 Multiple baseline design

Addressing the vaccine effects on the population overall ignores the difference between the direct effects on vaccinated children and indirect effects due to herd immunity. We estimated the age-specific effect of PCV introduction, however our model assumed the effect of the intervention to be common to all age groups. A multiple baseline design [80], an

extension of the CITS design, could be applied to measure the lag of PCV effect across groups, or to jointly assess the effect of PCV7 and PCV13. This kind of design assumes that the intervention is introduced in different groups at different times, with a different subset acting either as intervention or control groups at each time. This is similar to a stepped wedge cluster randomized trial, but typically does not involve randomization [123].

9.2.4 Indirect effect of reduced S.Pneumoniae circulation

We included counts of other respiratory bacterial infections, assuming no interference. However, future work might be needed to investigate whether, in reducing carriage of vaccine-type pneumococci, PCVs are creating an ecological niche favoring colonization by alternative respiratory pathogens such as *Staphylococcus aureus*, *Haemophilus influenzae*, and *Moraxella catarrhalis*.

9.3 Concluding remarks

This thesis focussed on identifying, extending and applying statistical models that had been previously employed in the analysis of routinely collected infectious disease data, to address their limitations.

Multiple modelling approaches can be explored when aiming to estimate disease burden or to evaluate effectiveness of interventions based on public health surveillance data. The standard approach in this field is to employ regression-based methods, while novel extensions of methods from the much-richer field of time-series are proposed here. Such time-series methods had only partially been exploited to date.

We have demonstrated here their usefulness and improved flexibility over the standard regression-based approaches in a novel setting: they allow modelling time dependence; and reduce the number of assumptions regarding seasonality and relationships between multiple time series.

Lastly, in applying these novel methods we have provided relevant estimates of influenza contribution to LRTI, pneumococcal vaccine impact and COVID-19 excess mortality, and clarified the contributing role of risk factors and heterogeneity across age groups in terms of

LRTI burden.

References

[1] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

[2] Abdullahi, O., Karani, A., Tigoi, C. C., Mugo, D., Kungu, S., Wanjiru, E., Jomo, J., Musyimi, R., Lipsitch, M., and Scott, J. A. G. (2012). The prevalence and risk factors for pneumococcal colonization of the nasopharynx among children in kilifi district, kenya. *PloS one*, 7(2):e30787.

[3] Aberth, J. (2016). *The Black Death: the great mortality of 1348-1350: a brief history with documents*. Springer.

[4] Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical methods in medical research*, 11(2):91–115.

[5] Auranen, K., Mehtälä, J., Tanskanen, A., and S Kaltoft, M. (2010). Between-strain competition in acquisition and clearance of pneumococcal carriage—epidemiologic evidence from a longitudinal study of day-care children. *American journal of epidemiology*, 171(2):169–176.

[6] Baker, P. J. (1992). T cell regulation of the antibody response to bacterial polysaccharide antigens: an examination of some general characteristics and their implications. *Journal of Infectious Diseases*, 165(Supplement_1):S44–S48.

[7] Balsells, E., Guillot, L., Nair, H., and Kyaw, M. H. (2017). Serotype distribution of streptococcus pneumoniae causing invasive disease in children in the post-pcv era: A systematic review and meta-analysis. *PloS one*, 12(5):e0177113.

[8] Baltrusaitis, K., Noddin, K., Nguyen, C., Crawley, A., Brownstein, J. S., and White, L. F. (2018). Evaluation of approaches that adjust for biases in participatory surveillance systems. *Online Journal of Public Health Informatics*, 10(1).

[9] Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

[10] Black, S., Shinefield, H., Fireman, B., Lewis, E., Ray, P., Hansen, J. R., Elvin, L., Ensor, K. M., Hackell, J., Siber, G., et al. (2000). Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. *The Pediatric infectious disease journal*, 19(3):187–195.

[11] Bonell, C. P., Hargreaves, J., Cousens, S., Ross, D., Hayes, R., Petticrew, M., and Kirkwood, B. (2011). Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions. *Journal of Epidemiology & Community Health*, 65(7):582–587.

[12] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (1976). *Time series analysis: forecasting and control*. John Wiley & Sons.

[13] Bracher, J. and Held, L. (2017). Periodically stationary multivariate non-gaussian autoregressive models. *arXiv preprint arXiv:1707.04635*.

[14] Bradley, E. and Kantz, H. (2015). Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097610.

[15] Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L., et al. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274.

[16] Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., and Rubin, G. J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet*.

[17] Casdagli, M., Eubank, S., Farmer, J. D., and Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, 51(1-3):52–98.

[18] Cawkwell, G. (2006). *Thucydides and the Peloponnesian war*. Routledge.

[19] CDC (2008). Public health surveillance- the best weapon to avert epidemics. Available from http://docshare.tips/dcpp-surveillance$_5$88*be*5*fcb*6*d*87*f*06598*b*46*dc.html*.

[20] Chatfield, C. (2016). *The analysis of time series: an introduction*. CRC press.

[21] Cherazard, R., Epstein, M., Doan, T.-L., Salim, T., Bharti, S., and Smith, M. A. (2017). Antimicrobial resistant streptococcus pneumoniae: prevalence, mechanisms, and clinical implications. *American journal of therapeutics*, 24(3):e361–e369.

[22] Chertow, D. S. and Memoli, M. J. (2013). Bacterial coinfection in influenza: a grand rounds review. *Jama*, 309(3):275–282.

[23] Chittaganpitch, M., Waicharoen, S., Yingyong, T., Praphasiri, P., Sangkitporn, S., Olsen, S. J., and Lindblade, K. A. (2018). Viral etiologies of influenza-like illness and severe acute respiratory infections in thailand. *Influenza and other respiratory viruses*, 12(4):482–489.

[24] Choi, B. C., Morrison, H., Wong, T., Wu, J., and Yan, Y.-P. (2007). Bringing chronic disease epidemiology and infectious disease epidemiology back together. *Journal of Epidemiology & Community Health*, 61(9):802–802.

[25] Choi, K. and Thacker, S. B. (1981). An evaluation of influenza mortality surveillance, 1962–1979 i. time series forecasts of expected pneumonia and influenza deaths. *American journal of epidemiology*, 113(3):215–226.

[26] Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., and Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96(5):1174–1181.

[27] Clifford, R. E., Smith, J., Tillett, H. E., and Wherry, P. J. (1977). Excess mortality associated with influenza in england and wales. *International journal of epidemiology*, 6(2):115–128.

[28] Cobey, S. and Baskerville, E. B. (2016). Limits to causal inference with state-space reconstruction for infectious disease. *PloS one*, 11(12).

[29] Cohen, R. (2006). Approaches to reduce antibiotic resistance in the community. *The Pediatric infectious disease journal*, 25(10):977–980.

[30] Collaborators, G. . L. R. I. et al. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet. Infectious diseases*, 18(11):1191.

[Columbia University] Columbia University. Difference-in-difference estimation.

[32] Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical methods in medical research*, 11(2):167–182.

[33] Cooper, B. S., Kotirum, S., Kulpeng, W., Praditsitthikorn, N., Chittaganpitch, M., Limmathurotsakul, D., Day, N. P., Coker, R., Teerawattananon, Y., and Meeyai, A. (2015). Mortality attributable to seasonal influenza a and b infections in thailand, 2005–2009: a longitudinal study. *American journal of epidemiology*, 181(11):898–907.

[34] Correa, A., Hinton, W., McGovern, A., van Vlymen, J., Yonova, I., Jones, S., and de Lusignan, S. (2016). Royal college of general practitioners research and surveillance centre (rcgp rsc) sentinel network: a cohort profile. *BMJ open*, 6(4):e011092.

[35] Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., and Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 93–115.

[36] Cox, D. R. and Miller, H. D. (1977). *The theory of stochastic processes*, volume 134. CRC Press.

[37] Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Ogilvie, D., Petticrew, M., Reeves, B., Sutton, M., et al. (2012). Using natural experiments to evaluate population health interventions: new medical research council guidance. *J Epidemiol Community Health*, 66(12):1182–1186.

[38] Crawford, D. H. (2007). *Deadly companions: How microbes shaped our history*. OUP Oxford.

[39] Cromer, D., van Hoek, A. J., Jit, M., Edmunds, W. J., Fleming, D., and Miller, E. (2014). The burden of influenza in england by age and clinical risk group: a statistical analysis to inform vaccine policy. *Journal of Infection*, 68(4):363–371.

[40] Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

[41] Dagan, R., Fraser, D., Givon, N., and Yagupsky, P. (1999). Carriage of resistant pneumococci by children in southern israel and impact of conjugate vaccines on carriage. *Clinical microbiology and infection*, 5:4S29–4S37.

[42] Danino, D., Givon-Lavi, N., Ben-Shimol, S., Greenberg, D., and Dagan, R. (2018). Understanding the evolution of antibiotic-nonsusceptible pneumococcal nasopharyngeal colonization following pneumococcal conjugate vaccine implementation in young children. *Clinical Infectious Diseases*.

[43] Davis, B. M., Aiello, A. E., Dawid, S., Rohani, P., Shrestha, S., and Foxman, B. (2012). Influenza and community-acquired pneumonia interactions: the impact of order and time of infection on population patterns. *American journal of epidemiology*, page kwr402.

[44] De Angelis, D., Presanis, A. M., Birrell, P. J., Tomba, G. S., and House, T. (2015). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10:83–87.

[45] De Rosa, S., Spaccarotella, C., Basso, C., Calabrò, M. P., Curcio, A., Filardi, P. P., Mancone, M., Mercuro, G., Muscoli, S., Nodari, S., et al. (2020). Reduction of hospitalizations for myocardial infarction in italy in the covid-19 era. *European Heart Journal*.

[46] Deyle, E. R., Maher, M. C., Hernandez, R. D., Basu, S., and Sugihara, G. (2016). Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences*, page 201607747.

[47] Dochez, A. and Gillespie, L. (1913). A biologic classification of pneumococci by means of immunity reactions. *Journal of the American Medical Association*, 61(10):727–732.

[48] Dowell, S. F. (2001). Seasonal variation in host susceptibility and cycles of certain infectious diseases. *Emerging infectious diseases*, 7(3):369.

[49] Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

[ECDC] ECDC. Graphs and maps.

[51] England, P. H. (2014). Standards for microbiology investigations (smi). Available from https://www.gov.uk/government/collections/standards-for-microbiology-investigations-smi.

[52] Engle, R. (2001). Garch 101: The use of arch/garch models in applied econometrics. *Journal of economic perspectives*, 15(4):157–168.

[53] Farré, L., Fasani, F., and Mueller, H. (2018). Feeling useless: the effect of unemployment on mental health in the great recession. *IZA Journal of Labor Economics*, 7(1):8.

[54] Farrington, C., Andrews, N. J., Beale, A., and Catchpole, M. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(3):547–563.

[55] Farrington, C., Kanaan, M., and Gay, N. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4(2):279–295.

[56] Fears, J. R. (2004). The plague under marcus aurelius and the decline and fall of the roman empire.

[57] Feikin, D. R., Kagucia, E. W., Loo, J. D., Link-Gelles, R., Puhan, M. A., Cherian, T., Levine, O. S., Whitney, C. G., O'Brien, K. L., Moore, M. R., et al. (2013). Serotype-specific changes in invasive pneumococcal disease after pneumococcal conjugate vaccine introduction: a pooled analysis of multiple surveillance sites. *PLoS medicine*, 10(9):e1001517.

[58] Feldman, C. and Anderson, R. (2014). Current and new generation pneumococcal vaccines. *Journal of Infection*, 69(4):309–325.

[59] for Disease Control, C., (CDC, P., et al. (2010). Licensure of a 13-valent pneumococcal conjugate vaccine (pcv13) and recommendations for use among children-advisory committee on immunization practices (acip), 2010. *MMWR. Morbidity and mortality weekly report*, 59(9):258.

[60] Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416.

[61] Gasparrini, A. (2014). Modeling exposure–lag–response associations with distributed lag non-linear models. *Statistics in medicine*, 33(5):881–899.

[62] Gay, N., Andrews, N., Trotter, C., and Edmunds, W. (2003). Estimating deaths due to influenza and respiratory syncytial virus—reply. *JAMA*, 289(19):2499–2502.

[63] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

[64] Geno, K. A., Gilbert, G. L., Song, J. Y., Skovsted, I. C., Klugman, K. P., Jones, C., Konradsen, H. B., and Nahm, M. H. (2015). Pneumococcal capsules and their types: past, present, and future. *Clinical microbiology reviews*, 28(3):871–899.

[65] George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

[66] George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

[67] Ghislandi, S., Muttarak, R., Sauerberg, M., and Scotti, B. (2020). News from the front: Estimation of excess mortality and life expectancy in the major epicenters of the covid-19 pandemic in italy. *medRxiv*.

[68] Gilca, R., De Serres, G., Skowronski, D., Boivin, G., and Buckeridge, D. L. (2009). The need for validation of statistical methods for estimating respiratory virus–attributable hospitalization. *American journal of epidemiology*, 170(7):925–936.

[69] Goettler, D., Streng, A., Kemmling, D., Schoen, C., von Kries, R., Rose, M., van der Linden, M., and Liese, J. (2020). Increase in streptococcus pneumoniae serotype 3 associated parapneumonic pleural effusion/empyema after the introduction of pcv13 in germany. *Vaccine*, 38(3):570–577.

[70] Goldstein, E., Viboud, C., Charu, V., and Lipsitch, M. (2012). Improving the estimation of influenza-related mortality over a seasonal baseline. *Epidemiology (Cambridge, Mass.)*, 23(6):829.

[71] Golyandina, N. and Korobeynikov, A. (2014). Basic singular spectrum analysis and forecasting with r. *Computational Statistics & Data Analysis*, 71:934–954.

[72] Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. CRC press.

[73] Graham, A. L., Allen, J. E., and Read, A. F. (2005). Evolutionary causes and consequences of immunopathology. *Annu. Rev. Ecol. Evol. Syst.*, 36:373–397.

[74] Granat, S. M., Mia, Z., Ollgren, J., Herva, E., Das, M., Piirainen, L., Auranen, K., and Mäkelä, P. H. (2007). Longitudinal study on pneumococcal carriage during the first year of life in bangladesh. *The Pediatric infectious disease journal*, 26(4):319–324.

[75] Granger, C. W. (1969). Testing for causality and feedback. *Econometrica*, 37(3):424–438.

[76] Gruger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics*, pages 595–605.

[77] Hanquet, G., Krizova, P., Valentiner-Branth, P., Ladhani, S. N., Nuorti, J. P., Lepoutre, A., Mereckiene, J., Knol, M., Winje, B. A., Ciruela, P., et al. (2019). Effect of childhood pneumococcal conjugate vaccination on invasive disease in older adults of 10 european countries: implications for adult vaccination. *Thorax*, 74(5):473–482.

[78] Harris, T. E. (2002). *The theory of branching processes*. Courier Corporation.

[79] Hassani, H. (2007). Singular spectrum analysis: methodology and comparison.

[80] Hawkins, N. G., Sanson-Fisher, R. W., Shakeshaft, A., D'Este, C., and Green, L. W. (2007). The multiple baseline design for evaluating population-based research. *American journal of preventive medicine*, 33(2):162–168.

[81] Hays, J. N. (2005). *Epidemics and pandemics: their impacts on human history*. Abc-clio.

[82] He, M. M. (2016). Driving through the great recession: Why does motor vehicle fatality decrease when the economy slows down? *Social Science & Medicine*, 155:1–11.

[83] Healy, M. (1983). A simple method for monitoring routine statistics. *The Statistician*, pages 347–349.

[84] Hegger, R., Kantz, H., and Schreiber, T. (1999). Practical implementation of nonlinear time series methods: The tisean package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435.

[85] Held, L., Hofmann, M., Höhle, M., and Schmid, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics*, 7(3):422–437.

[86] Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical modelling*, 5(3):187–199.

[87] Held, L., Meyer, S., and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: the 13th armitage lecture. *Statistics in medicine*, 36(22):3443–3460.

[88] Hellard, E., Fouchet, D., Vavre, F., and Pontier, D. (2015). Parasite–parasite interactions in the wild: How to detect them? *Trends in parasitology*, 31(12):640–652.

[89] Hendriks, W., Boshuizen, H., Dekkers, A., Knol, M., Donker, G. A., van der Ende, A., and Korthals Altes, H. (2017). Temporal cross-correlation between influenza-like illnesses and invasive pneumococcal disease in the netherlands. *Influenza and Other Respiratory Viruses*.

[90] Henrichsen, J. (1995). Six newly recognized types of streptococcus pneumoniae. *Journal of clinical microbiology*, 33(10):2759.

[91] Hill, P. C., Cheung, Y. B., Akisanya, A., Sankareh, K., Lahai, G., Greenwood, B. M., and Adegbola, R. A. (2008). Nasopharyngeal carriage of streptococcus pneumoniae in gambian infants: a longitudinal study. *Clinical infectious diseases*, 46(6):807–814.

[92] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

[93] Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, 20(1):5–10.

[94] Hopkins, D. R. (1980). Ramses v: earliest know victim?

[95] Hougaard, P. (1999). Multi-state models: a review. *Lifetime data analysis*, 5(3):239–264.

[96] Hubert, B., Watier, L., Garnerin, P., and Richardson, S. (1992). Meningococcal disease and influenza-like syndrome: a new approach to an old question. *Journal of Infectious Diseases*, 166(3):542–545.

[97] Huffaker, R., Bittelli, M., and Rosa, R. (2017). *Nonlinear time series analysis with R*. Oxford University Press.

[98] Hulten, K. G. (2018). The changing epidemiology of pneumococcal diseases. *The Lancet Infectious Diseases*, 18(9):929–930.

[99] Ihekweazu, C. A., Dance, D., Pebody, R., George, R., Smith, M., Waight, P., Christensen, H., Cartwright, K., and Stuart, J. (2008). Trends in incidence of pneumococcal disease before introduction of conjugate vaccine: South west england, 1996–2005. *Epidemiology & Infection*, 136(8):1096–1102.

[100] Imai, C. and Hashizume, M. (2015). A systematic review of methodology: time series regression analysis for environmental factors and infectious diseases. *Tropical medicine and health*, 43(1):1–9.

[101] Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

[Ipsos Mori] Ipsos Mori. The health foundation covid-19 survey.

[103] Izurieta, H. S., Thompson, W. W., Kramarz, P., Shay, D. K., Davis, R. L., DeStefano, F., Black, S., Shinefield, H., and Fukuda, K. (2000). Influenza and the rates of hospitalization for respiratory disease among infants and young children. *New England Journal of Medicine*, 342(4):232–239.

[104] Jackson, M., Peterson, D., Nelson, J., Greene, S., Jacobsen, S., Belongia, E., Baxter, R., and Jackson, L. A. (2015). Using winter 2009–2010 to assess the accuracy of methods which estimate influenza-related morbidity and mortality. *Epidemiology and infection*, 143(11):2399–2407.

[105] Jackson, M. L. (2009). Confounding by season in ecologic studies of seasonal exposures and outcomes: examples from estimates of mortality due to influenza. *Annals of Epidemiology*, 19(10):681–691.

[106] Janeway, C. A., Travers, P., Walport, M., Shlomchik, M., et al. (1996). *Immunobiology: the immune system in health and disease*, volume 7. Current Biology London.

[107] Kalbfleisch, J. and Lawless, J. F. (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80(392):863–871.

[108] Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis*, volume 7. Cambridge university press.

[109] Karppinen, S., Teräsjärvi, J., Auranen, K., Schuez-Havupalo, L., Siira, L., He, Q., Waris, M., and Peltola, V. (2017). Acquisition and transmission of streptococcus pneumoniae are facilitated during rhinovirus infection in families with children. *American journal of respiratory and critical care medicine*, 196(9):1172–1180.

[110] Kennel, M. B., Brown, R., and Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403.

[111] Klugman, K. P. (2001). Efficacy of pneumococcal conjugate vaccines and their effect on carriage and antimicrobial resistance. *The Lancet infectious diseases*, 1(2):85–91.

[112] Kuster, S. P., Tuite, A. R., Kwong, J. C., McGeer, A., Fisman, D. N., Network, T. I. B. D., et al. (2011). Evaluation of coseasonality of influenza and invasive pneumococcal disease: results from prospective surveillance. *PLoS Med*, 8(6):e1001042.

[113] Ladhani, S. N., Collins, S., Djennad, A., Sheppard, C. L., Borrow, R., Fry, N. K., Andrews, N. J., Miller, E., and Ramsay, M. E. (2018). Rapid increase in non-vaccine serotypes causing invasive pneumococcal disease in england and wales, 2000–17: a prospective national observational cohort study. *The Lancet Infectious Diseases*, 18(4):441–451.

[114] Launes, C., de Sevilla, M.-F., Selva, L., Garcia-Garcia, J.-J., Pallares, R., and Muñoz-Almagro, C. (2012). Viral coinfection in children less than five years old with invasive pneumococcal disease. *The Pediatric infectious disease journal*, 31(6):650–653.

[115] Lawless, J. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542.

[116] Lewis, D. (1973). Counterfactuals and comparative possibility. In *Ifs*, pages 57–85. Springer.

[117] Lewnard, J. A. and Hanage, W. P. (2019). Making sense of differences in pneumococcal serotype replacement. *The Lancet Infectious Diseases*.

[118] Linden, A. and Yarnold, P. R. (2016). Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*, 22(6):855–859.

[119] Linley, E., Bell, A., Gritzfeld, J., and Borrow, R. (2019). Should pneumococcal serotype 3 be included in serotype-specific immunoassays? *Vaccines*, 7(1):4.

[120] Lipsitch, M., Abdullahi, O., D'Amour, A., Xie, W., Weinberger, D. M., Tchetgen, E. T., and Scott, J. A. G. (2012). Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in kenya with a markov transition model. *Epidemiology (Cambridge, Mass.)*, 23(4):510.

[121] Littman, R. J. (2009). The plague of athens: epidemiology and paleopathology. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine*, 76(5):456–467.

[122] Loi, M. and Rodrigues, M. (2012). A note on the impact evaluation of public policies: the counterfactual analysis.

[123] Lopez Bernal, J., Cummins, S., and Gasparrini, A. (2018). The use of controls in interrupted time series studies of public health interventions. *International journal of epidemiology*, 47(6):2082–2093.

[124] Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141.

[125] Lui, K.-J. and Kendal, A. P. (1987). Impact of influenza epidemics on mortality in the united states from october 1972 to may 1985. *American journal of public health*, 77(6):712–716.

[126] Madhav, N., Oppenheim, B., Gallivan, M., Mulembakani, P., Rubin, E., and Wolfe, N. (2017). Pandemics: risks, impacts, and mitigation. In *Disease Control Priorities: Improving Health and Reducing Poverty. 3rd edition*. The International Bank for Reconstruction and Development/The World Bank.

[127] Makela, P. (2008). Bj. history of pneumococcal immunization siber gr kk, makela ph.

[128] Marks, G. and Beatty, W. K. (1976). *Epidemics*. Scribner.

[129] Matias, G., Taylor, R. J., Haguinet, F., Schuck-Paim, C., Lustig, R. L., and Fleming, D. M. (2016). Modelling estimates of age-specific influenza-related hospitalisation and mortality in the united kingdom. *BMC public health*, 16(1):481.

[130] Mbelle, N., Huebner, R. E., Wasas, A. D., Kimura, A., Chang, I., and Klugman, K. P. (1999). Immunogenicity and impact on nasopharyngeal carriage of a nonavalent pneumococcal conjugate vaccine. *The Journal of infectious diseases*, 180(4):1171–1176.

[131] McCullers, J. A. (2006). Insights into the interaction between influenza virus and pneumococcus. *Clinical microbiology reviews*, 19(3):571–582.

[132] McCullers, J. A. (2014). The co-pathogenesis of influenza viruses with bacteria in the lung. *Nature Reviews Microbiology*, 12(4):252–262.

[133] McIsaac, D. I., Abdulla, K., Yang, H., Sundaresan, S., Doering, P., Vaswani, S. G., Thavorn, K., and Forster, A. J. (2017). Association of delay of urgent or emergency surgery with mortality and use of health care resources: a propensity score–matched observational cohort study. *Cmaj*, 189(27):E905–E912.

[134] McLaughlin, J. M., Jiang, Q., Gessner, B. D., Swerdlow, D. L., Sings, H. L., Isturiz, R. E., and Jodar, L. (2019). Pneumococcal conjugate vaccine against serotype 3 pneumococcal pneumonia in adults: A systematic review and pooled analysis. *Vaccine*.

[135] McNeill, W. H. (1993). Patterns of disease emergence in history. *Emerging viruses*, pages 29–36.

[136] Mehr, S. and Wood, N. (2012). Streptococcus pneumoniae–a review of carriage, infection, serotype replacement and vaccination. *Paediatric respiratory reviews*, 13(4):258–264.

[137] Mehtälä, J., Antonio, M., Kaltoft, M. S., O'Brien, K. L., and Auranen, K. (2013). Competition between streptococcus pneumoniae strains: implications for vaccine-induced replacement in colonization and disease. *Epidemiology*, pages 522–529.

[138] Melegaro, A., Choi, Y., Pebody, R., and Gay, N. (2007). Pneumococcal carriage in united kingdom families: estimating serotype-specific transmission parameters from longitudinal data. *American journal of epidemiology*, 166(2):228–235.

[139] Meyer, S., Held, L., et al. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3):1612–1639.

[140] Meyer, S., Held, L., and Höhle, M. (2016). hhh4: Endemic-epidemic modeling of areal count time series. *Journal of Statistical Software*.

[141] Millar, E. V., O'Brien, K. L., Zell, E. R., Bronsdon, M. A., Reid, R., and Santosham, M. (2009). Nasopharyngeal carriage of streptococcus pneumoniae in navajo and white mountain apache children before the introduction of pneumococcal conjugate vaccine. *The Pediatric infectious disease journal*, 28(8):711–716.

[142] Miller, E., Andrews, N. J., Waight, P. A., Slack, M. P., and George, R. C. (2011). Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in england and wales: an observational cohort study. *The Lancet infectious diseases*, 11(10):760–768.

[143] Mills, T. C. (1991). *Time series techniques for economists*. Cambridge University Press.

[144] Mina, M. J. and Klugman, K. P. (2014). The role of influenza in the severity and transmission of respiratory bacterial disease. *The Lancet Respiratory Medicine*, 2(9):750–763.

[145] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

[146] Mølbak, K. and Mazick, A. (2013). European monitoring of excess mortality for public health action (euromomo) kåre mølbak. *European Journal of Public Health*, 23(suppl_1).

[147] Montemurro, N. (2020). The emotional impact of covid-19: From medical staff to common people. *Brain, behavior, and immunity*.

[148] Morens, D. M., Taubenberger, J. K., and Fauci, A. S. (2008). Predominant role of bacterial pneumonia as a cause of death in pandemic influenza: implications for pandemic influenza preparedness. *Journal of Infectious Diseases*, 198(7):962–970.

[149] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3):e74.

[150] Murdoch, D. R. and Howie, S. R. (2018). The global burden of lower respiratory infections: making progress, but we need to do better. *The Lancet Infectious Diseases*, 18(11):1162–1163.

[151] Murdoch, D. R. and Jennings, L. C. (2009). Association of respiratory virus activity and environmental factors with the incidence of invasive pneumococcal disease. *Journal of Infection*, 58(1):37–46.

[152] NHS England. Ae attendances and emergency admissions may 2020 statistical commentary.

[153] NHS England. Cancer waiting times data.

[154] NHS England. Nhs inpatient admission and outpatient referrals and attendances.

[155] NHS England. Nhs referral to treatment (rtt) waiting times data.

[156] Nicoli, E. J., Trotter, C. L., Turner, K. M., Colijn, C., Waight, P., and Miller, E. (2013). Influenza and rsv make a modest contribution to invasive pneumococcal disease incidence in the uk. *Journal of Infection*, 66(6):512–520.

[157] Obaro, S. K., Adegbola, R., Banya, W., and Greenwood, B. (1996). Carriage of pneumococci after pneumococcal vaccination. *The Lancet*, 348(9022):271–272.

[158] O'Brien, K. L., Moulton, L. H., Reid, R., Weatherholtz, R., Oski, J., Brown, L., Kumar, G., Parkinson, A., Hu, D., Hackell, J., et al. (2003). Efficacy and safety of seven-valent conjugate pneumococcal vaccine in american indian children: group randomised trial. *The Lancet*, 362(9381):355–361.

[159] O'Brien, K. L., Wolfson, L. J., Watt, J. P., Henkle, E., Deloria-Knoll, M., McCall, N., Lee, E., Mulholland, K., Levine, O. S., Cherian, T., et al. (2009). Burden of disease caused by streptococcus pneumoniae in children younger than 5 years: global estimates. *The Lancet*, 374(9693):893–902.

[160] Olsen, L. F. and Schaffer, W. M. (1990). Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics. *Science*, 249(4968):499–504.

[161] ONS (2017). England population mid-year estimate, office for national statistics. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/timeserie

[162] Opatowski, L., Varon, E., Dupont, C., Temime, L., van der Werf, S., Gutmann, L., Boëlle, P.-Y., Watier, L., and Guillemot, D. (2013). Assessing pneumococcal meningitis association with viral respiratory infections and antibiotics: insights from statistical and mathematical models. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1764):20130519.

[163] Ouldali, N., Levy, C., Minodier, P., Morin, L., Biscardi, S., Aurel, M., Dubos, F., Dommergues, M. A., Mezgueldi, E., Levieux, K., et al. (2019). Long-term association of 13-valent pneumococcal conjugate vaccine implementation with rates of community-acquired pneumonia in children. *JAMA pediatrics*, 173(4):362–370.

[164] Park, R. E. and Mitchell, B. M. (1980). Estimating the autocorrelated error model with trended data. *Journal of Econometrics*, 13(2):185–201.

[165] Parker, D. E., Legg, T. P., and Folland, C. K. (1992). A new daily central england temperature series, 1772–1991. *International Journal of Climatology*, 12(4):317–342.

[166] Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in medicine*, 27(29):6250.

[167] Peter, Ď. and Silvia, P. (2012). Arima vs. arimax–which approach is better to analyze and forecast macroeconomic time series. In *Proceedings of 30th International Conference Mathematical Methods in Economics. Karviná, Czech Republic*, pages 136–140.

[168] Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer.

[169] Petticrew, M., Cummins, S., Ferrell, C., Findlay, A., Higgins, C., Hoy, C., Kearns, A., and Sparks, L. (2005). Natural experiments: an underused tool for public health? *Public health*, 119(9):751–757.

[170] Pick, H., Daniel, P., Rodrigo, C., Bewick, T., Ashton, D., Lawrence, H., Baskaran, V., Edwards-Pritchard, R. C., Sheppard, C., Eletu, S. D., et al. (2020). Pneumococcal serotype trends, surveillance and risk factors in uk adult pneumonia, 2013–18. *Thorax*, 75(1):38–49.

[171] Pilishvili, T., Lexau, C., Farley, M. M., Hadler, J., Harrison, L. H., Bennett, N. M., Reingold, A., Thomas, A., Schaffner, W., Craig, A. S., et al. (2010). Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine. *The Journal of infectious diseases*, 201(1):32–41.

[172] Politis, D. N. (2001). Resampling time series with seasonal components. In *Frontiers in data mining and bioinformatics: Proceedings of the 33rd symposium on the interface of computing science and statistics*, pages 13–17.

[173] Poole, J. and Holladay, A. J. (1979). Thucydides and the plague of athens. *The Classical Quarterly*, 29(2):282–300.

[174] Presanis, A. M., Pebody, R. G., Birrell, P. J., Tom, B. D., Green, H. K., Durnall, H., Fleming, D., De Angelis, D., et al. (2014). Synthesising evidence to estimate pandemic (2009) a/h1n1 influenza severity in 2009–2011. *The Annals of Applied Statistics*, 8(4):2378–2403.

[175] Priestley, M. (1981). *Spectral Analysis and Time Series. Vol. 2. Multivariate Series Prediction and Control*.

[176] Raoult, D., Mouffok, N., Bitam, I., Piarroux, R., and Drancourt, M. (2013). Plague: history and contemporary analysis. *Journal of Infection*, 66(1):18–26.

[177] Richardson, S., Stücker, I., and Hémon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, 16(1):111–120.

[178] Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern epidemiology*. Lippincott Williams & Wilkins.

[179] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

[180] Ryan, A. M., Krinsky, S., Kontopantelis, E., and Doran, T. (2016). Long-term evidence for the effect of pay-for-performance in primary care on mortality in the uk: a population study. *The Lancet*, 388(10041):268–274.

[181] Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81.

[182] Saunders, J., Lundberg, R., Braga, A. A., Ridgeway, G., and Miles, J. (2015). A synthetic control approach to evaluating place-based crime interventions. *Journal of Quantitative Criminology*, 31(3):413–434.

[183] Schaffer, W. M. and Kot, M. (1985). Nearly one dimensional dynamics in an epidemic. *Journal of Theoretical Biology*, 112(2):403–427.

[184] Schanzer, D., Tam, T., Langley, J., and Winchester, B. (2007). Influenza-attributable deaths, canada 1990–1999. *Epidemiology and Infection*, 135(07):1109–1116.

[185] Serfling, R. E. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6):494.

[186] Shadish, W. R., Cook, T. D., Campbell, D. T., et al. (2002). *Experimental and quasi-experimental designs for generalized causal inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell.* Boston: Houghton Mifflin,.

[187] Shaman, J. and Kohn, M. (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248.

[188] Shiri, T., McCarthy, N. D., and Petrou, S. (2019). The impact of childhood pneumococcal vaccination on hospital admissions in england: a whole population observational study. *BMC infectious diseases*, 19(1):510.

[189] Shrestha, S., Foxman, B., Berus, J., Van Panhuis, W. G., Steiner, C., Viboud, C., and Rohani, P. (2015). The role of influenza in the epidemiology of pneumonia. *Scientific reports*, 5.

[190] Shrestha, S., Foxman, B., Weinberger, D. M., Steiner, C., Viboud, C., and Rohani, P. (2013). Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. *Science translational medicine*, 5(191):191ra84–191ra84.

[191] Simell, B., Auranen, K., Käyhty, H., Goldblatt, D., Dagan, R., and O'Brien, K. L. (2012). The fundamental link between pneumococcal carriage and disease. *Expert review of vaccines*, 11(7):841–855.

[192] Simonsen, L., Blackwelder, W., Reichert, T., and Miller, M. (2003). Estimating deaths due to influenza and respiratory syncytial virus. *JAMA*, 289(19):2499–2502.

[193] Simonsen, L., Clarke, M. J., Williamson, G. D., Stroup, D. F., Arden, N. H., and Schonberger, L. B. (1997). The impact of influenza epidemics on mortality: introducing a severity index. *American journal of public health*, 87(12):1944–1950.

[194] Simonsen, L., Fukuda, K., Schonberger, L. B., and Cox, N. J. (2000). The impact of influenza epidemics on hospitalizations. *Journal of Infectious Diseases*, 181(3):831–837.

[195] Stensballe, L. G., Hjuler, T., Andersen, A., Kaltoft, M., Ravn, H., Aaby, P., and Simoes, E. A. (2008). Hospitalization for respiratory syncytial virus infection and invasive pneumococcal disease in danish children aged< 2 years: a population-based cohort study. *Clinical infectious diseases*, 46(8):1165–1171.

[196] Stilianakis, N. I. and Drossinos, Y. (2010). Dynamics of infectious disease transmission by inhalable respiratory droplets. *Journal of the Royal Society Interface*, 7(50):1355–1366.

[197] Strogatz, S. (2001). Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (studies in nonlinearity).

[198] Sugihara, G., Grenfell, B. T., and May, R. M. (1990). Distinguishing error from chaos in ecological time series. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 330(1257):235–251.

[199] Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *science*, 338(6106):496–500.

[200] Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer.

[201] Taubenberger, J. K. and Morens, D. M. (2006). 1918 influenza: the mother of all pandemics. *Rev Biomed*, 17:69–79.

[202] Theiler, J. and Eubank, S. (1993). Don't bleach chaotic data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 3(4):771–782.

[203] Thompson, W. W., Shay, D. K., Weintraub, E., Brammer, L., Cox, N., Anderson, L. J., and Fukuda, K. (2003). Mortality associated with influenza and respiratory syncytial virus in the united states. *Jama*, 289(2):179–186.

[204] Thompson, W. W., Weintraub, E., Dhankhar, P., Cheng, P.-Y., Brammer, L., Meltzer, M. I., Bresee, J. S., and Shay, D. K. (2009). Estimates of us influenza-associated deaths made using four different methods. *Influenza and other respiratory viruses*, 3(1):37–49.

[205] Thorrington, D., Andrews, N., Stowe, J., Miller, E., and van Hoek, A. J. (2018). Elucidating the impact of the pneumococcal conjugate vaccine programme on pneumonia, sepsis and otitis media hospital admissions in england using a composite control. *BMC medicine*, 16(1):13.

[206] Tidd, C., Olsen, L., and Schaffer, W. M. (1993). The case for chaos in childhood epidemics. ii. predicting historical epidemics from mathematical models. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 254(1341):257–273.

[207] Trotter, C. L., Gay, N. J., and Edmunds, W. J. (2005). Dynamic models of meningococcal carriage, disease, and the impact of serogroup c conjugate vaccination. *American journal of epidemiology*, 162(1):89–100.

[208] Trotter, C. L., Waight, P., Andrews, N. J., Slack, M., Efstratiou, A., George, R., and Miller, E. (2010). Epidemiology of invasive pneumococcal disease in the pre-conjugate vaccine era: England and wales, 1996–2006. *Journal of Infection*, 60(3):200–208.

[209] Trotter Jr, Y., Dunn, F. L., Drachman, R. H., Henderson, D. A., Pizzi, M., Langmuir, A. D., et al. (1959). Asian influenza in the united states, 1957-1958. *American journal of hygiene*, 70(1):34–50.

[210] Tsonis, A. A., Deyle, E. R., May, R. M., Sugihara, G., Swanson, K., Verbeten, J. D., and Wang, G. (2015). Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proceedings of the National Academy of Sciences*, 112(11):3253–3256.

[211] Turner, P., Turner, C., Jankhot, A., Helen, N., Lee, S. J., Day, N. P., White, N. J., Nosten, F., and Goldblatt, D. (2012). A longitudinal study of streptococcus pneumoniae carriage in a cohort of infants and their mothers on the thailand-myanmar border. *PloS one*, 7(5):e38271.

[212] Turner, P., Turner, C. L., Watthanaworawit, W., Carrara, V. I., Kapella, B. K., Painter, J., Nosten, F. H., et al. (2010). Influenza in refugees on the thailand-myanmar border, may-october 2009. *Emerg Infect Dis*, 16(9):1366–1372.

[213] Unkel, S., Farrington, C., Garthwaite, P. H., Robertson, C., and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):49–82.

[214] Uusitalo, L., Lehikoinen, A., Helle, I., and Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*, 63:24–31.

[215] Vadlamudi, N. K., Chen, A., and Marra, F. (2018). Impact of the 13-valent pneumococcal conjugate vaccine among adults: A systematic review and meta-analysis. *Clinical Infectious Diseases*, 69(1):34–49.

[216] van der Linden, M. and Whitney, C. G. (2019). Herd protection or herding cats?

[217] van Hoek, A. J., Choi, Y. H., Trotter, C., Miller, E., and Jit, M. (2012). The cost-effectiveness of a 13-valent pneumococcal conjugate vaccination for infants in england. *Vaccine*, 30(50):7205–7213.

[218] van Hoek, A. J., Sheppard, C. L., Andrews, N. J., Waight, P. A., Slack, M. P., Harrison, T. G., Ladhani, S. N., and Miller, E. (2014). Pneumococcal carriage in children and adults two years after introduction of the thirteen valent pneumococcal conjugate vaccine in england. *Vaccine*, 32(34):4349–4355.

[219] Van Nes, E. H., Scheffer, M., Brovkin, V., Lenton, T. M., Ye, H., Deyle, E., and Sugihara, G. (2015). Causal feedbacks in climate change. *Nature Climate Change*, 5(5):445–448.

[220] Vandenbroucke, J. P., Rooda, H. E., and Beukers, H. (1991). Who made john snow a hero? *American Journal of Epidemiology*, 133(10):967–973.

[221] Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and therapeutics*, 40(4):277.

[222] Vernet, G., Saha, S., Satzke, C., Burgess, D., Alderson, M., Maisonneuve, J.-F., Beall, B., Steinhoff, M., and Klugman, K. (2011). Laboratory-based diagnosis of pneumococcal pneumonia: state of the art and unmet needs. *Clinical Microbiology and Infection*, 17:1–13.

[223] Vestjens, S. M., Sanders, E. A., Vlaminckx, B. J., de Melker, H. E., van der Ende, A., and Knol, M. J. (2019). Twelve years of pneumococcal conjugate vaccination in the netherlands: Impact on incidence and clinical outcomes of invasive pneumococcal disease. *Vaccine*, 37(43):6558–6565.

[224] Vynnycky, E. and White, R. (2010). *An introduction to infectious disease modelling*. OUP oxford.

[225] Wagner, A. K., Soumerai, S. B., Zhang, F., and Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics*, 27(4):299–309.

[226] Waight, P. A., Andrews, N. J., Ladhani, S. N., Sheppard, C. L., Slack, M. P., and Miller, E. (2015). Effect of the 13-valent pneumococcal conjugate vaccine on invasive pneumococcal disease in england and wales 4 years after its introduction: an observational cohort study. *The Lancet infectious diseases*, 15(5):535–543.

[227] Wang, X., Piao, S., Ciais, P., Friedlingstein, P., Myneni, R. B., Cox, P., Heimann, M., Miller, J., Peng, S., Wang, T., et al. (2014a). A two-fold increase of carbon cycle sensitivity to tropical temperature variations. *Nature*, 506(7487):212–215.

[228] Wang, X.-L., Wong, C.-M., Chan, K.-H., Chan, K.-P., Cao, P.-H., Peiris, J. M., and Yang, L. (2014b). Hospitalization risk of the 2009 h1n1 pandemic cases in hong kong. *BMC infectious diseases*, 14(1):32.

[229] Wang, Y., Yang, J., Chen, Y., De Maeyer, P., Li, Z., and Duan, W. (2018). Detecting the causal effect of soil moisture on precipitation using convergent cross mapping. *Scientific reports*, 8(1):1–8.

[230] Weinberger, D. M., Harboe, Z. B., Viboud, C., Krause, T. G., Miller, M., Mølbak, K., and Konradsen, H. B. (2014). Pneumococcal disease seasonality: incidence, severity and the role of influenza activity. *European Respiratory Journal*, 43(3):833–841.

[231] Weinberger, D. M., Klugman, K. P., Steiner, C. A., Simonsen, L., and Viboud, C. (2015). Association between respiratory syncytial virus activity and pneumococcal disease in infants: a time series analysis of us hospitalization data. *PLoS medicine*, 12(1):e1001776.

[232] Weinberger, D. M., Simonsen, L., Jordan, R., Steiner, C., Miller, M., and Viboud, C. (2011). Impact of the 2009 influenza pandemic on pneumococcal pneumonia hospitalizations in the united states. *Journal of Infectious Diseases*, 205(3):458–465.

[WHO] WHO. Measles cases spike globally due to gaps in vaccination coverage.

[234] WHO (2013). Child mortality estimates due to hib and pneumococcal infections. Available from https://www.who.int/immunization/monitoring$_s urveillance/burden/estimates/Pneumo_h ib/en/$.

[235] Wing, C., Simon, K., and Bello-Gomez, R. A. (2018). Designing difference in difference studies: best practices for public health policy research. *Annual review of public health*, 39.

[236] Yang, L., Chiu, S. S., Chan, K.-P., Chan, K.-H., Wong, W. H.-S., Peiris, J. M., and Wong, C.-M. (2011). Validation of statistical models for estimating hospitalization associated with influenza and other respiratory viruses. *PLoS One*, 6(3):e17882.

[237] Ye, H., Deyle, E. R., Gilarranz, L. J., and Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5:14750.

[238] Young, T. K. (2004). *Population health: concepts and methods*. Oxford University Press.

[239] Zhou, H., Thompson, W. W., Viboud, C. G., Ringholz, C. M., Cheng, P.-Y., Steiner, C., Abedi, G. R., Anderson, L. J., Brammer, L., and Shay, D. K. (2012). Hospitalizations associated with influenza and respiratory syncytial virus in the united states, 1993–2008. *Clinical infectious diseases*, 54(10):1427–1436.

Appendix A

Supplementary information to chapter 4

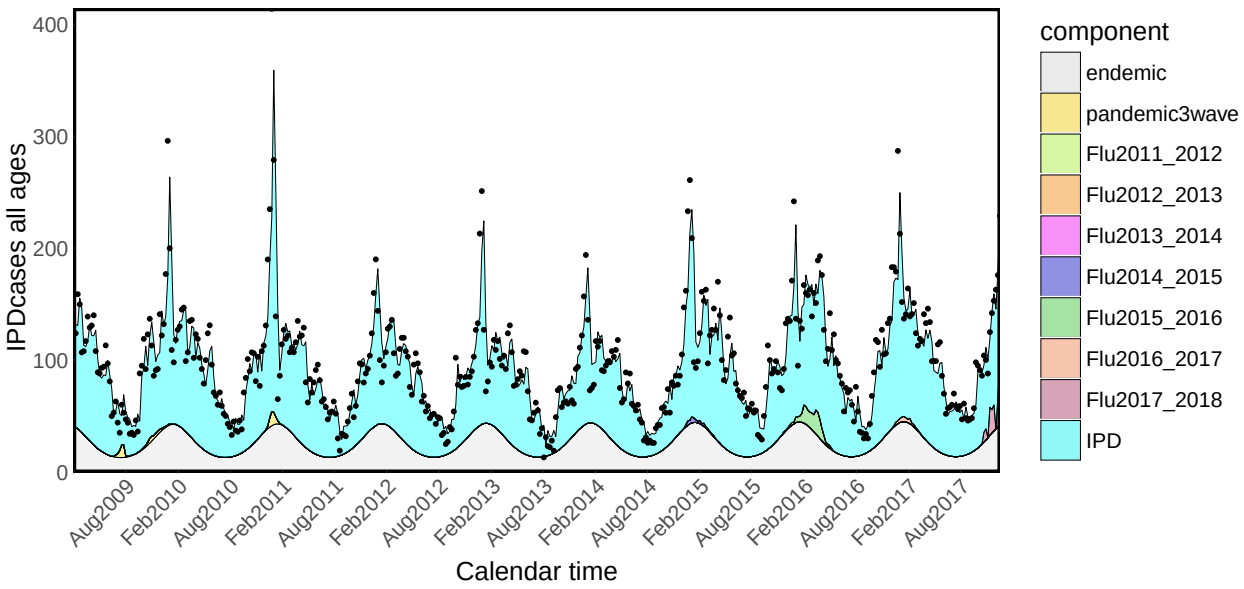


Fig. A.1 Fitted IPD values all ages

Age	α	γ	δ	$log(\psi)$	$log(\tau)$	$log(\theta)$	$log(\zeta)$	$log(\lambda)$	$log(\phi)$
<5	-2.369	-0.319	-0.062	2.524	-	-2.307	-4.000	2.178	1.225
5-14	-4.395	-0.367	-0.062	1.598	-3.245	-	-	2.282	1.308
15-44	-4.034	-0.477	-0.062	3.027	-1.627	-	-	3.725	4.094
45-64	-2.912	-0.315	-0.062	3.304	-2.156	2.215	2.754	3.158	3.646
65+	-2.046	-0.458	-0.062	3.341	-	2.504	2.571	3.097	4.337

Table A.1 Model K: Coefficient estimates for the age-specific model of IPD including Flu, rhinovirus and RSV

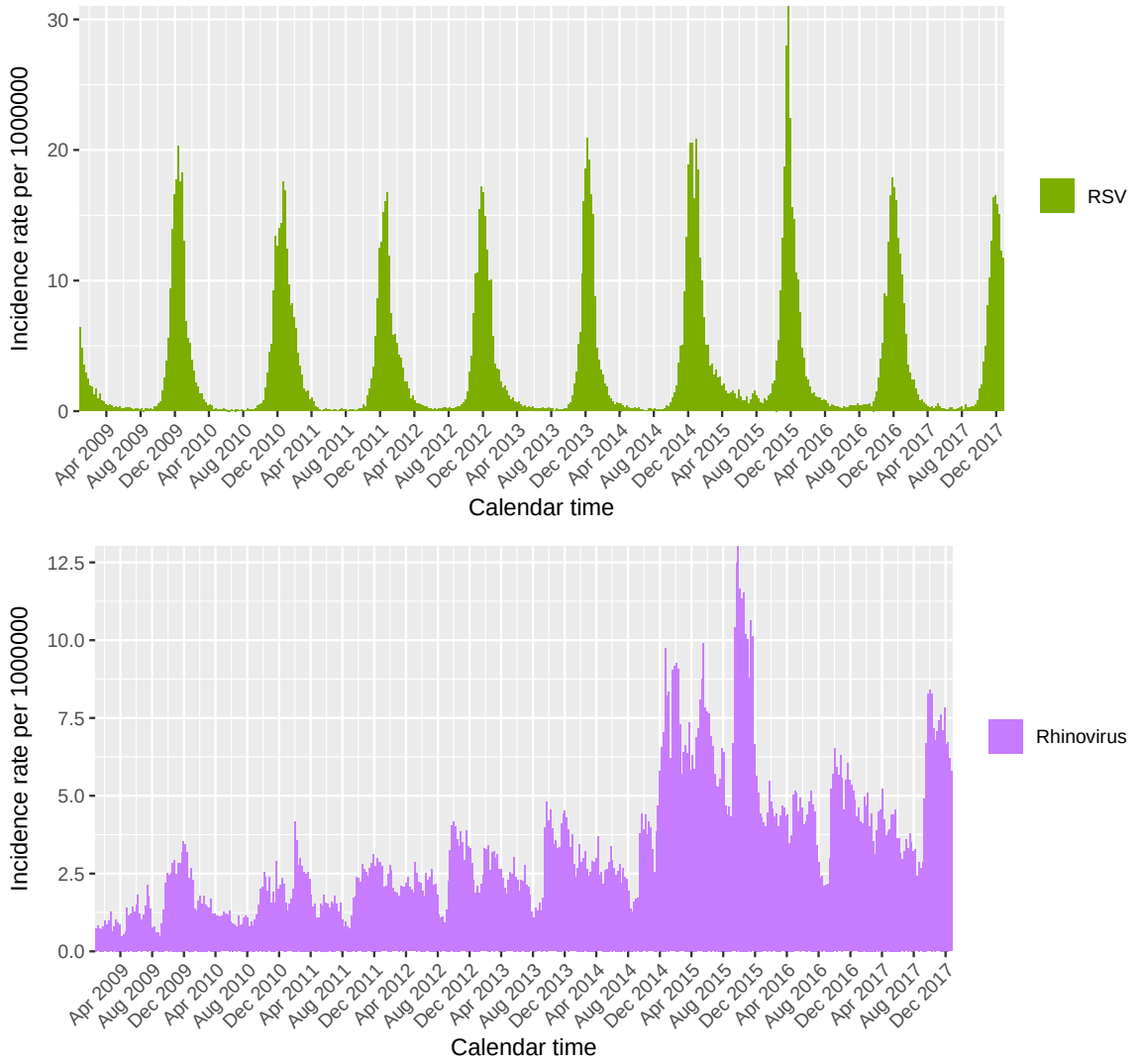


Fig. A.2 RSV and rhinovirus incidence rates

Age	α	γ	δ	$log(\psi)$	$log(\tau)$	$log(\theta)$	$log(\zeta)$	$log(\lambda)$	$log(\phi)$
< 5	0.016	0.017	0.003	0.039	-	0.311	1.140	0.011	0.016
$5-14$	0.014	0.009	0.003	0.072	0.100	-	-	0.052	0.025
$15-44$	0.006	0.006	0.003	0.022	0.066	-	-	0.007	0.001
$45-64$	0.014	0.012	0.003	0.018	0.671	0.040	0.051	0.006	0.001
$65+$	0.004	0.003	0.003	0.011	-	0.028	0.047	0.003	0.001

Table A.2 Model K: Coefficient standard errors for the age-specific model of IPD including Flu, rhinovirus and RSV

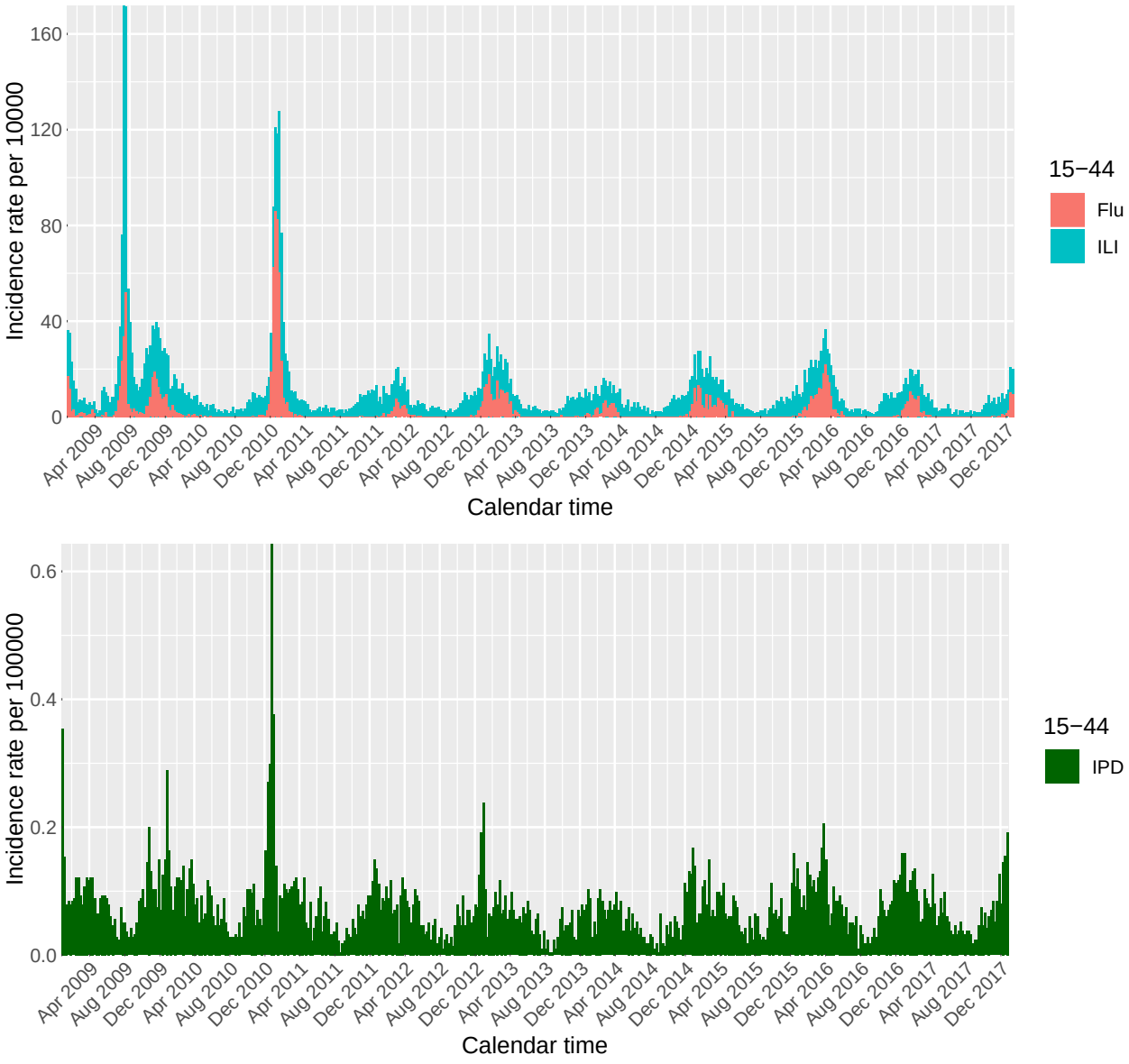


Fig. A.3 Observed counts: 15-44 years old

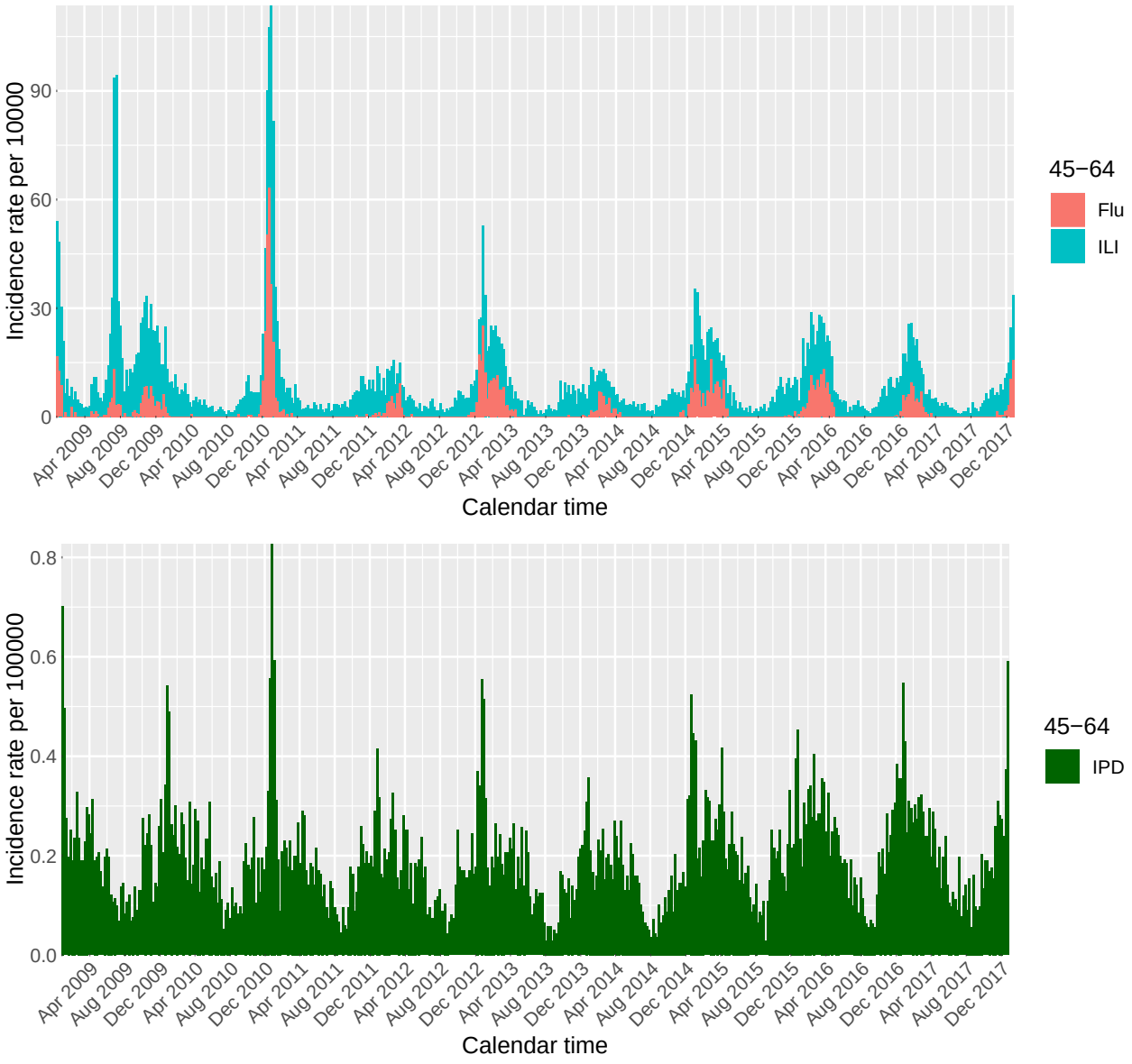


Fig. A.4 Observed counts: 45-64 years old

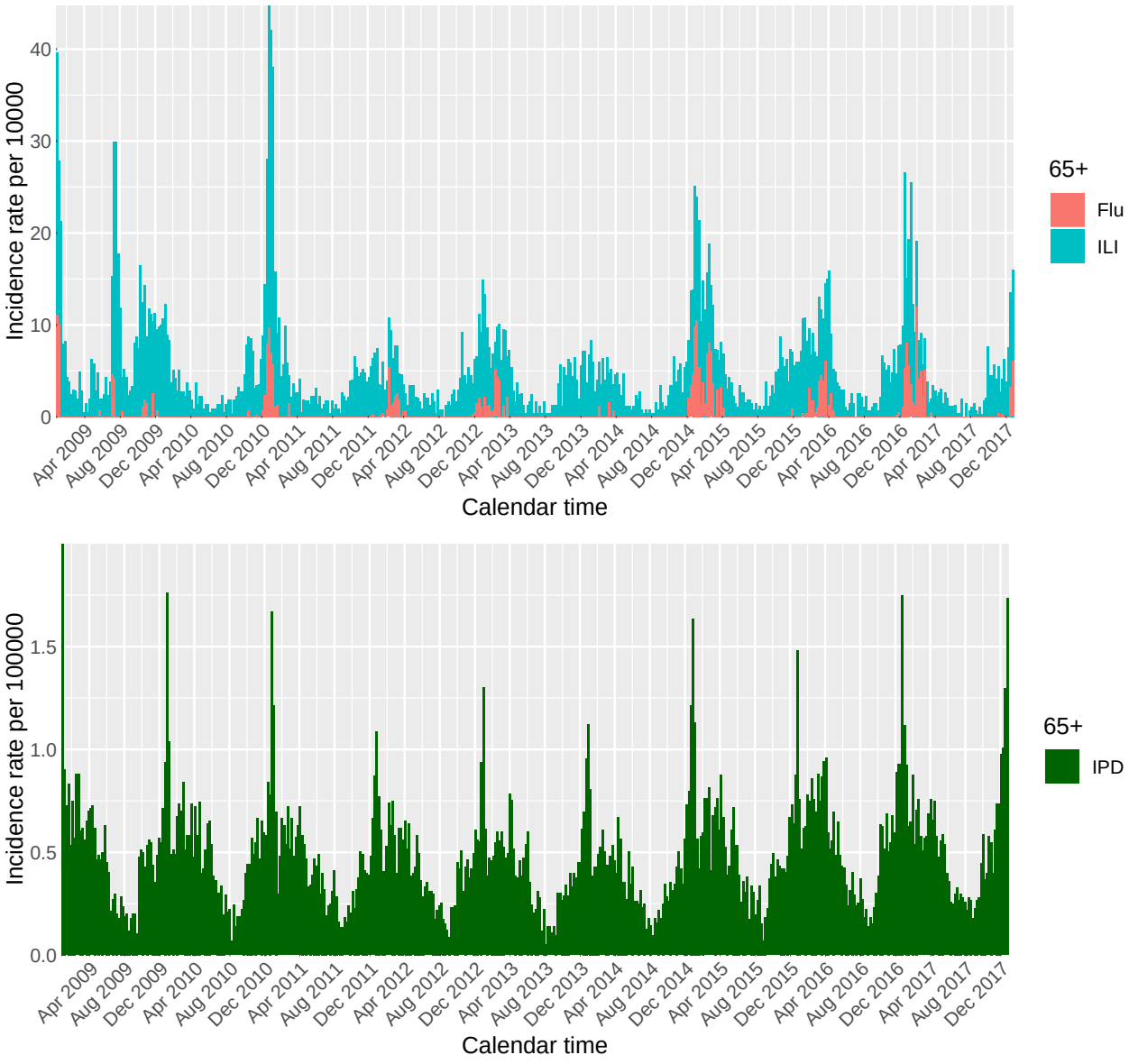


Fig. A.5 Observed counts: 65+ years old

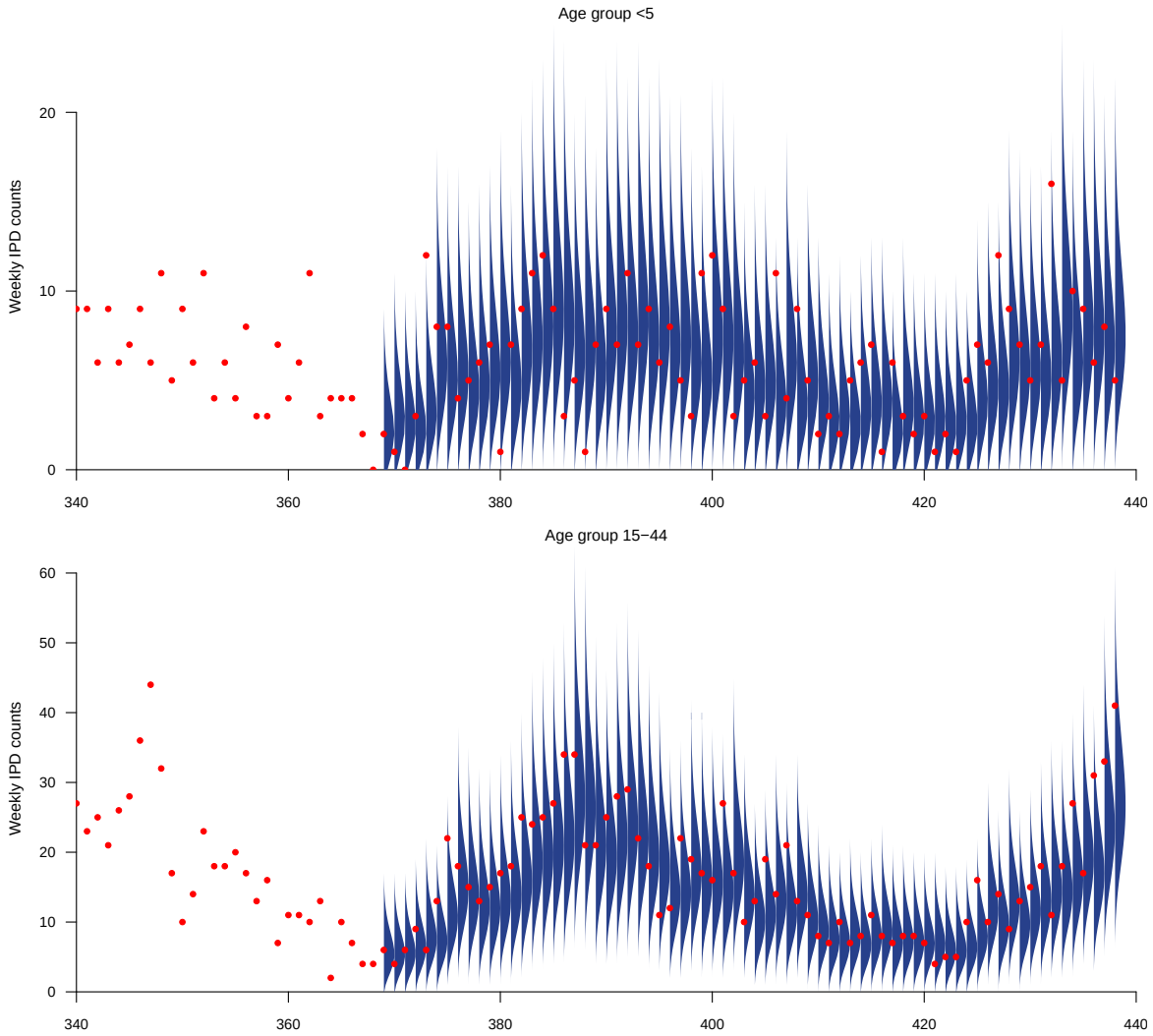


Fig. A.6 Model H: Predictive distribution for infants and young adults

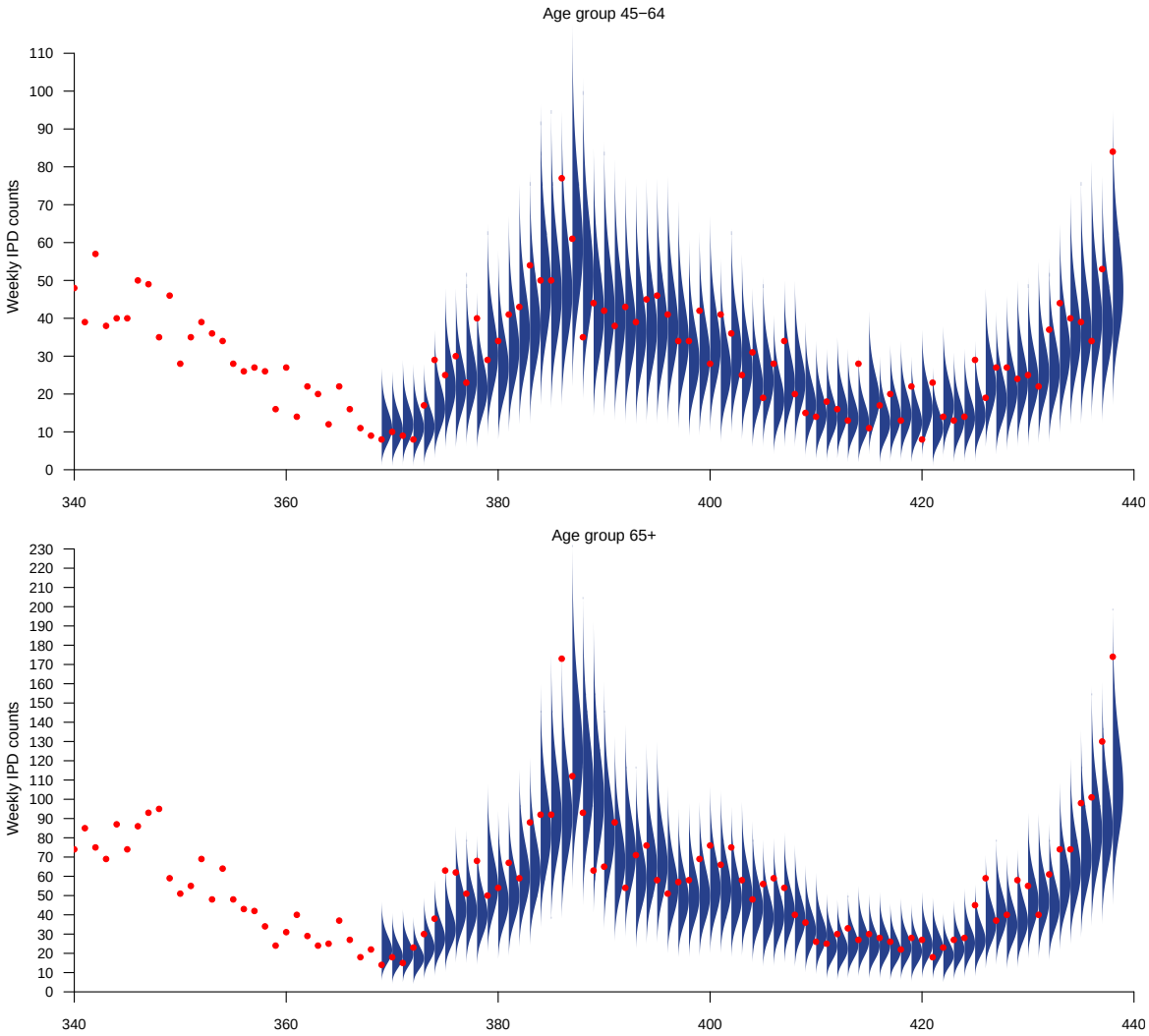


Fig. A.7 Model H: Predictive distribution for 45-64 and 65+

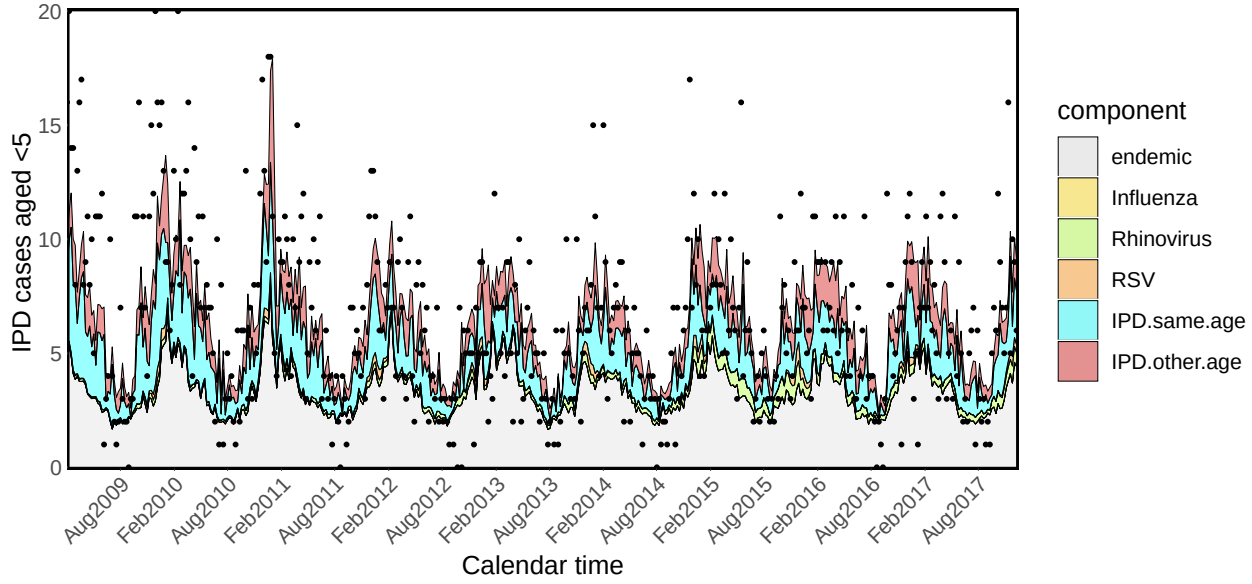


Fig. A.8 Model K: Fitted IPD values for infants

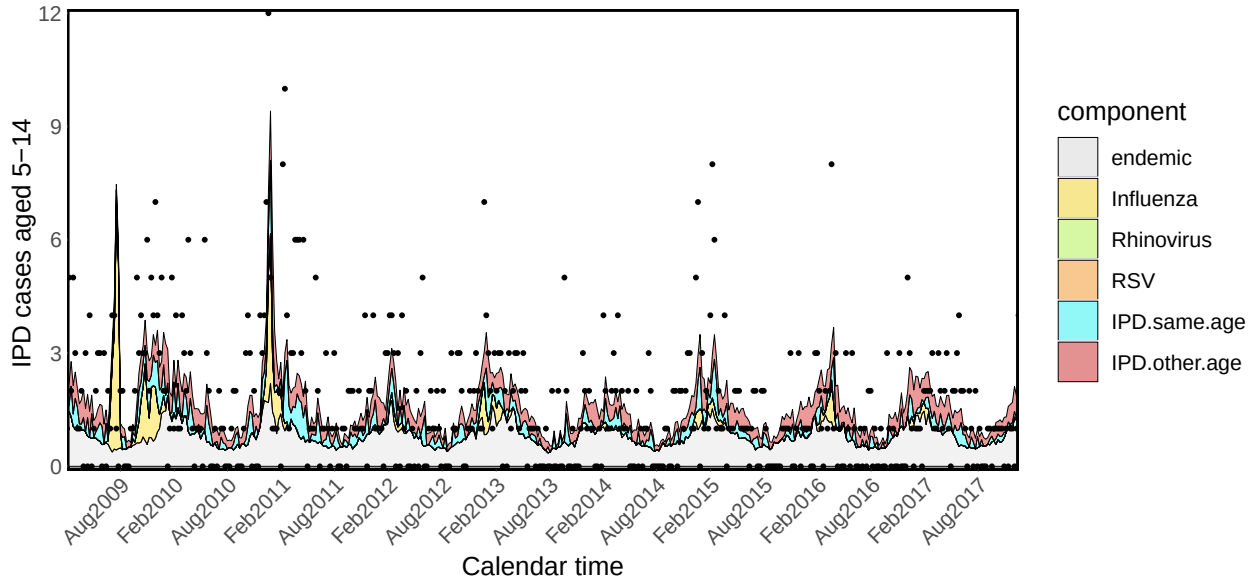


Fig. A.9 Model K: Fitted IPD values for school-age children

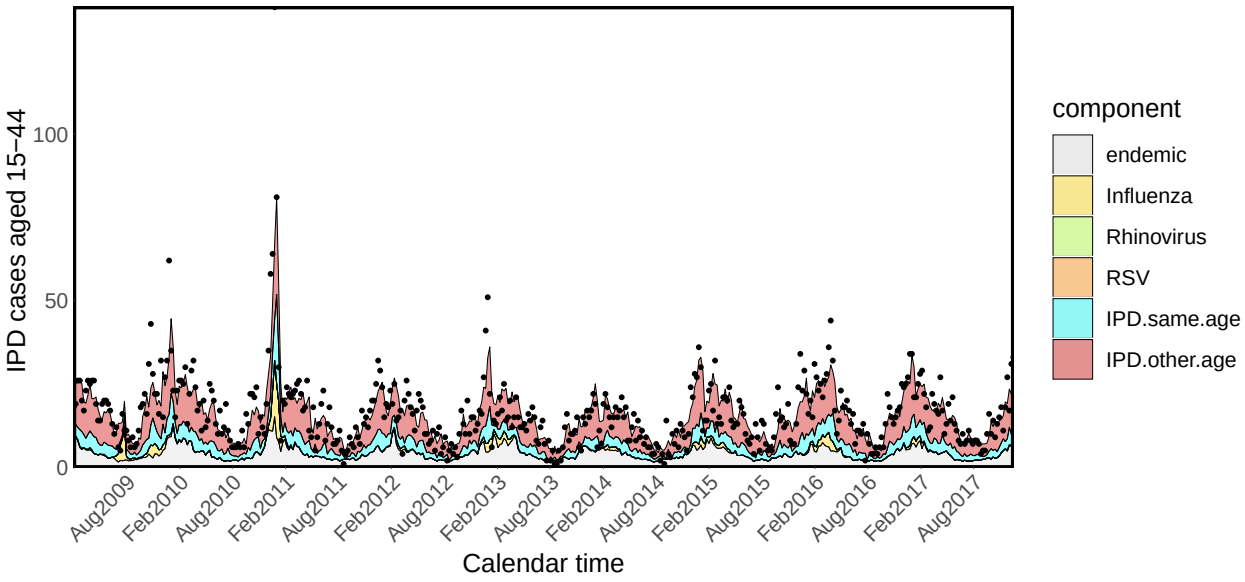


Fig. A.10 Model K: Fitted IPD values for young adults

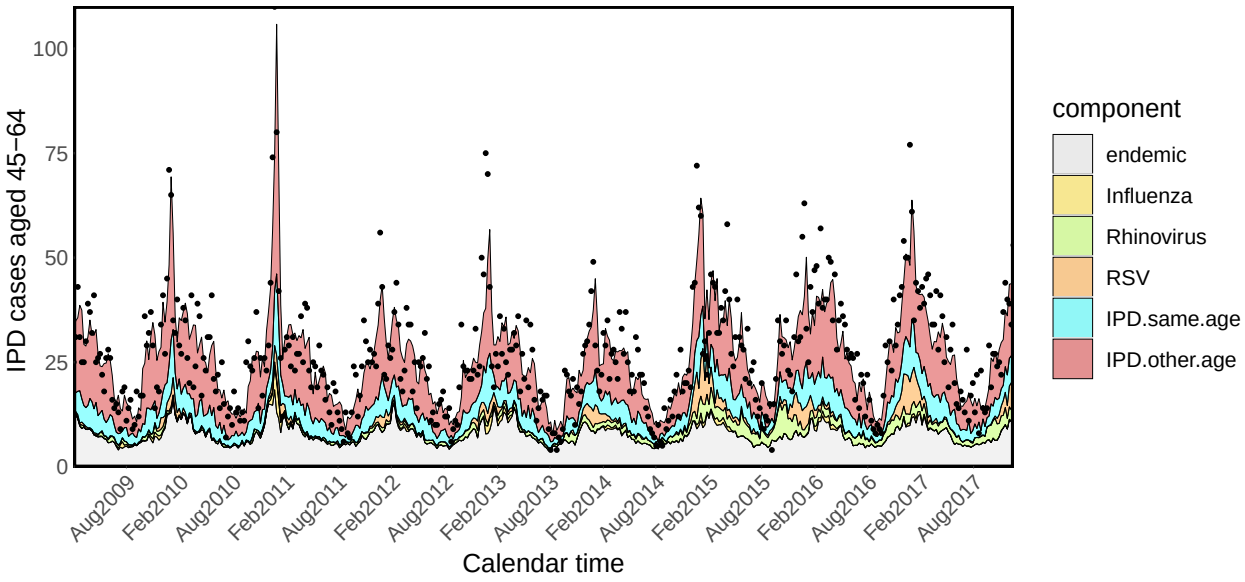


Fig. A.11 Model K: Fitted IPD values for the 45-64 age group

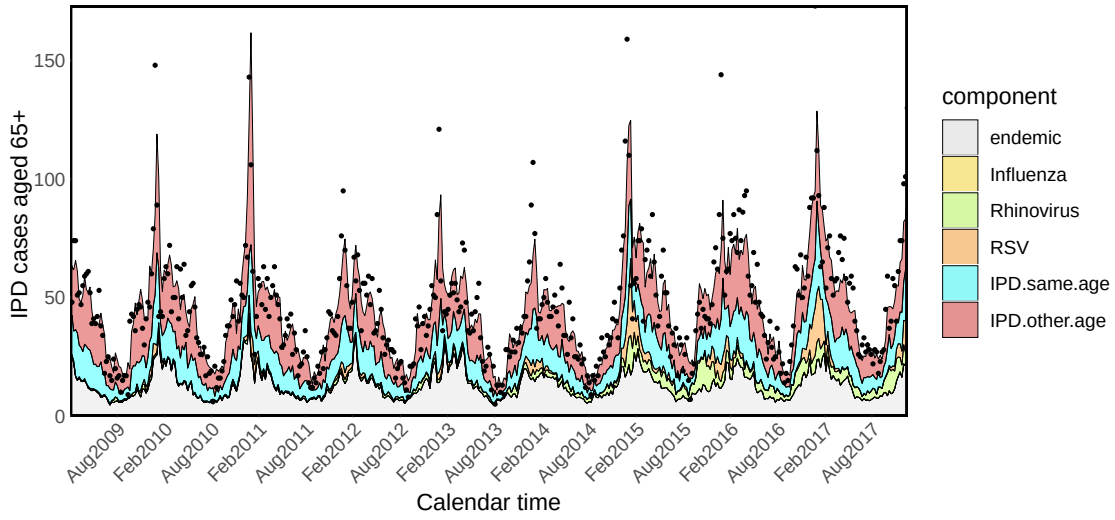


Fig. A.12 Model K: Fitted IPD values for the elderly

Appendix B

Supplementary information to chapter 6

B.1 Observed IPD incidence by age and PCV group

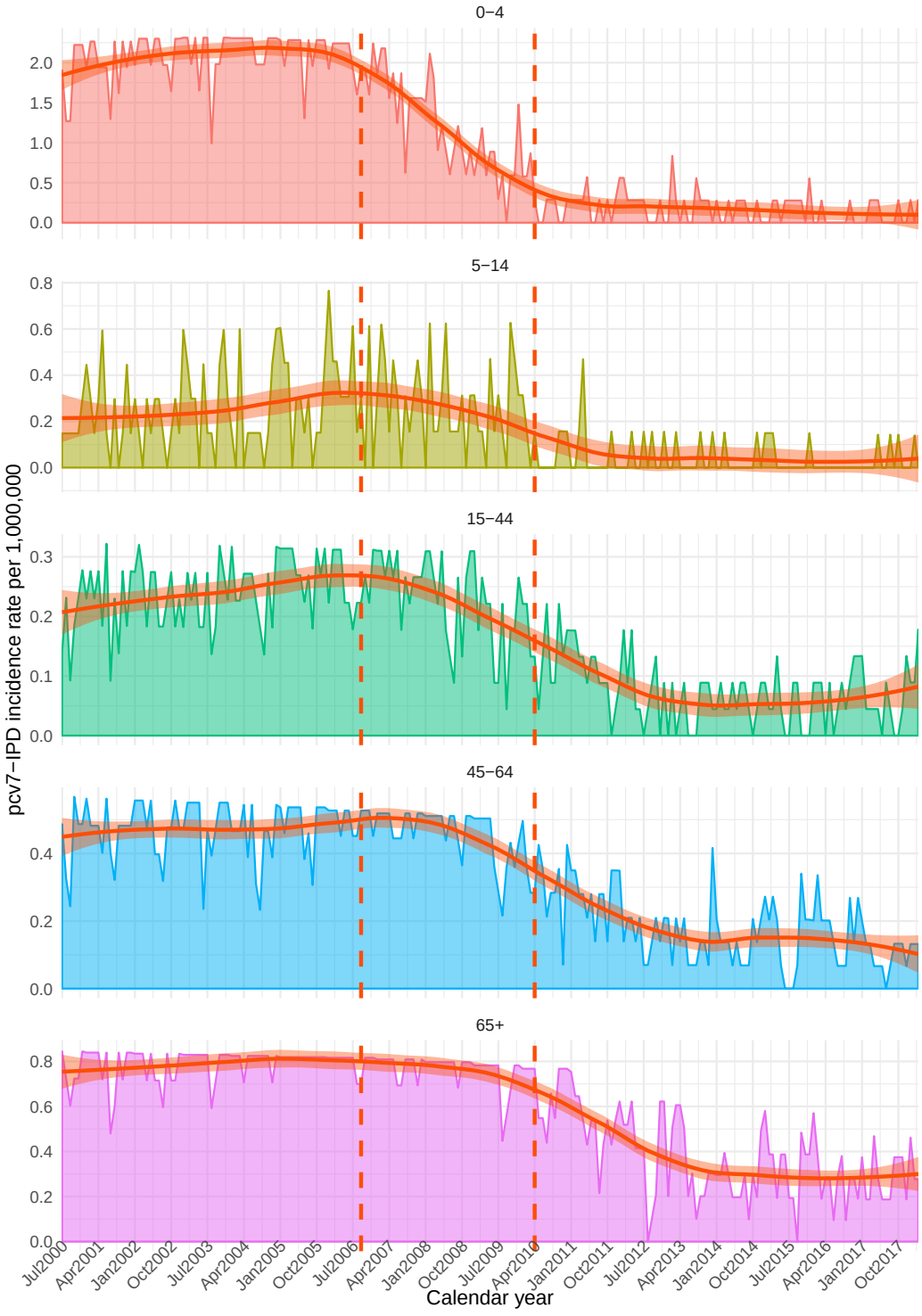


Fig. B.1 PCV7 serotypes, incidence rate per million residents (scales differ across panels)

.

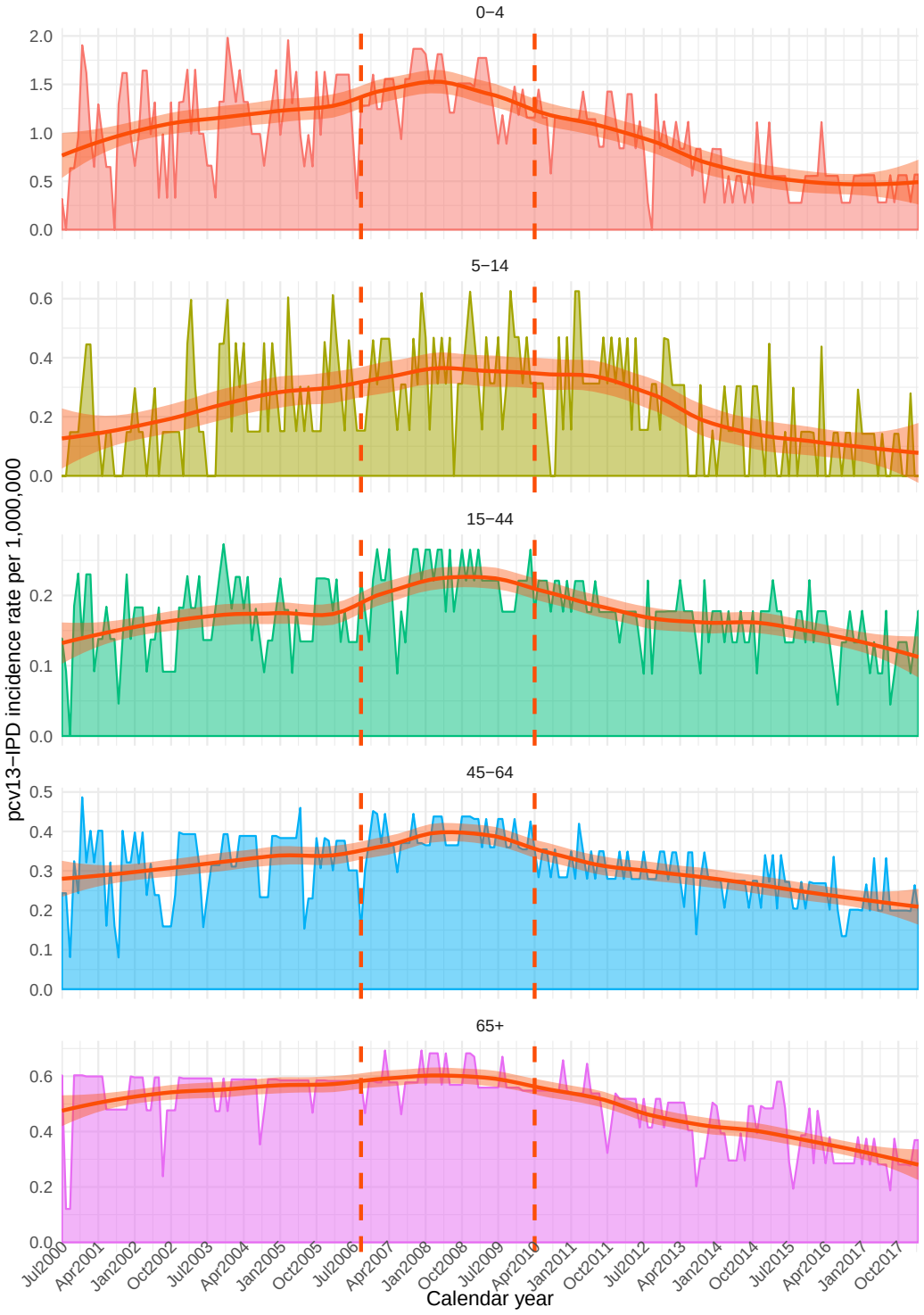


Fig. B.2 PCV13 serotypes, incidence rate per million residents (scales differ across panels)

.

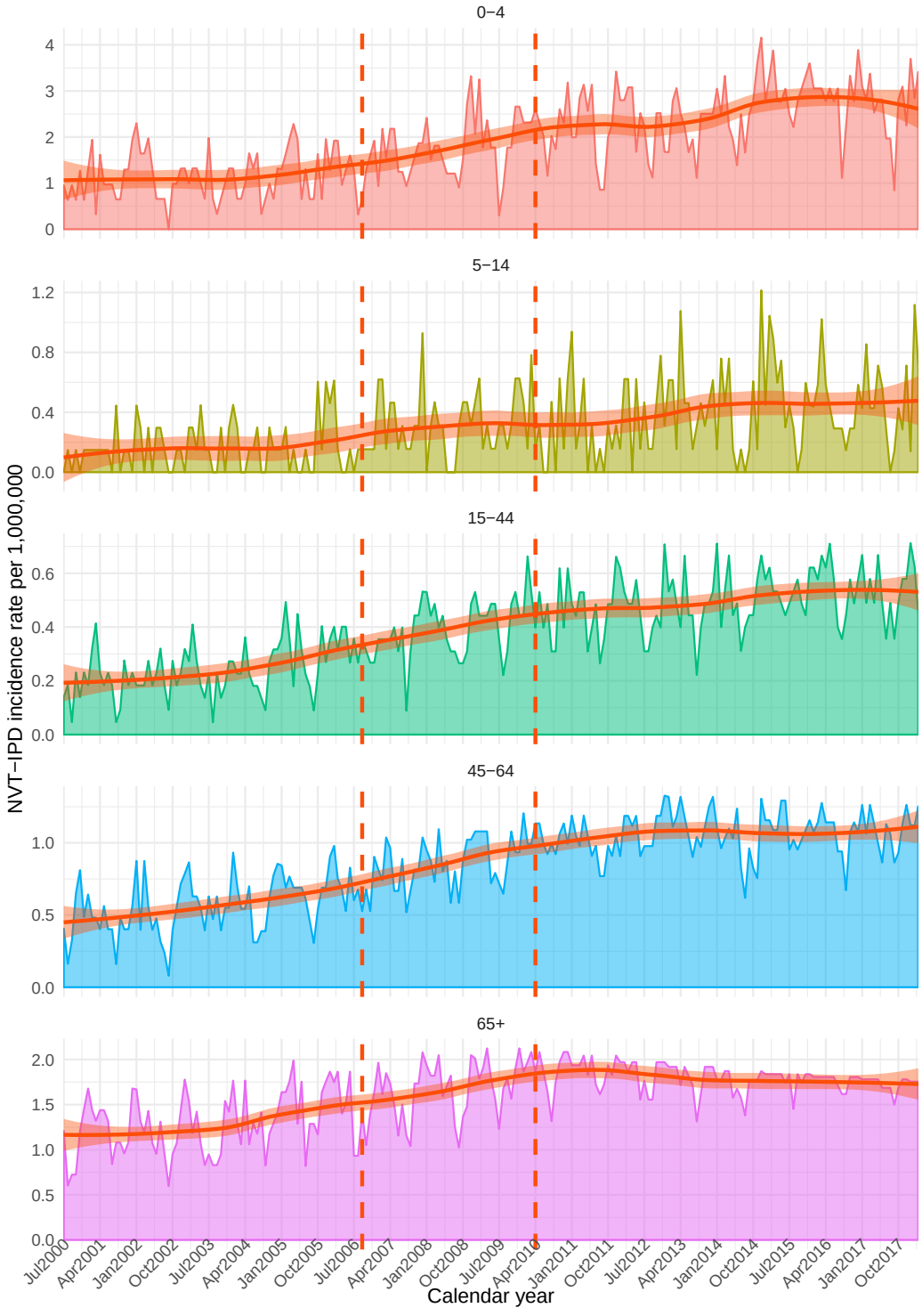


Fig. B.3 NV serotypes, incidence rate per million residents (scales differ across panels)

.

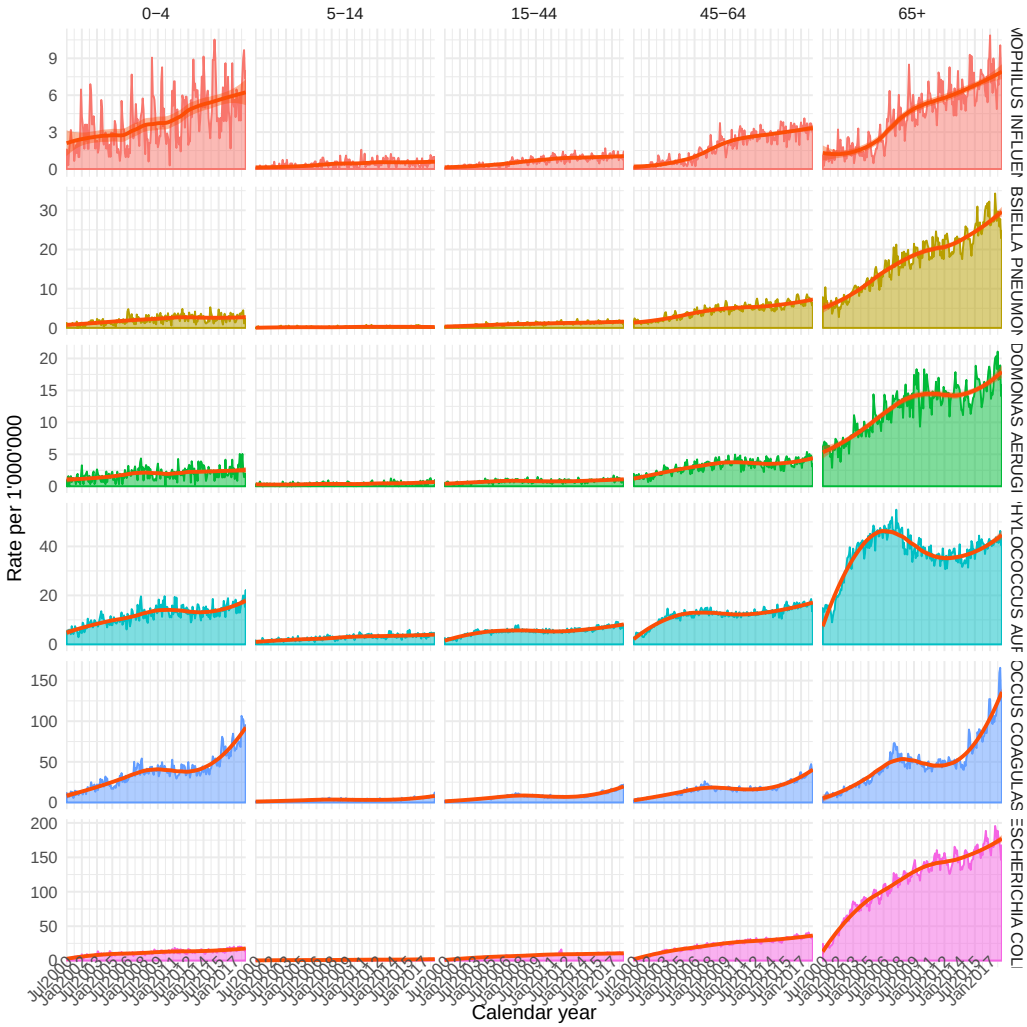


Fig. B.4 Control time series, incidence rate per million residents (scales differ across panels)

.

B.2 Sensitivity analysis for change of lag in ITS

The results presented in the main text assume that impact of intervention kicks in at one year lag from the policy implementation. Here we explore how using a two-years lag affects our conclusions. We can say that change of lag does not produce a significant change when modelling the impact of PCV overall, as shown in figure B.5. Similarly, when modelling PCV7-IPD and non-PCV7-IPD, we can see that the two-year lag slightly amplifies the intervention effect, as for B.6, but effects have the same direction and magnitude. Finally, results don't change considerably also when modelling PCV13-IPD and nonPCV13-IPD, as shown in B.7. Hence, we are overall confident that our results are robust with respect to the

chosen lag.

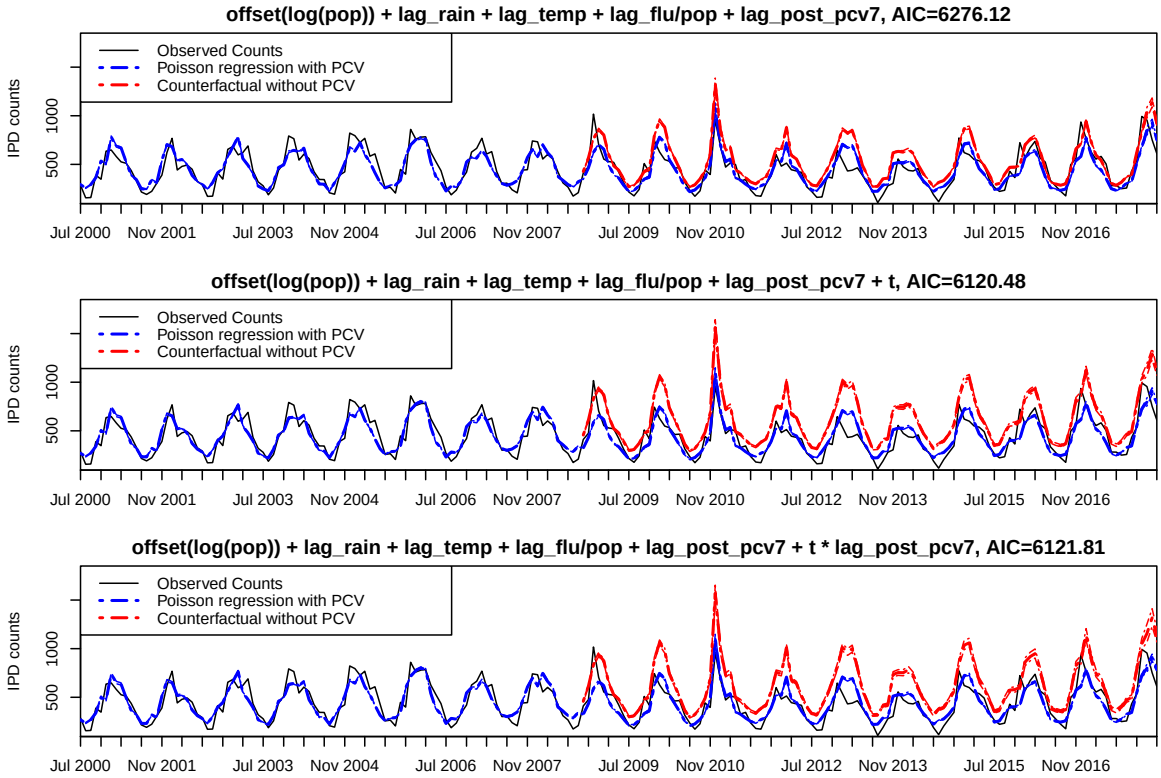


Fig. B.5 Fitted IPD counts based on three ITS models

.

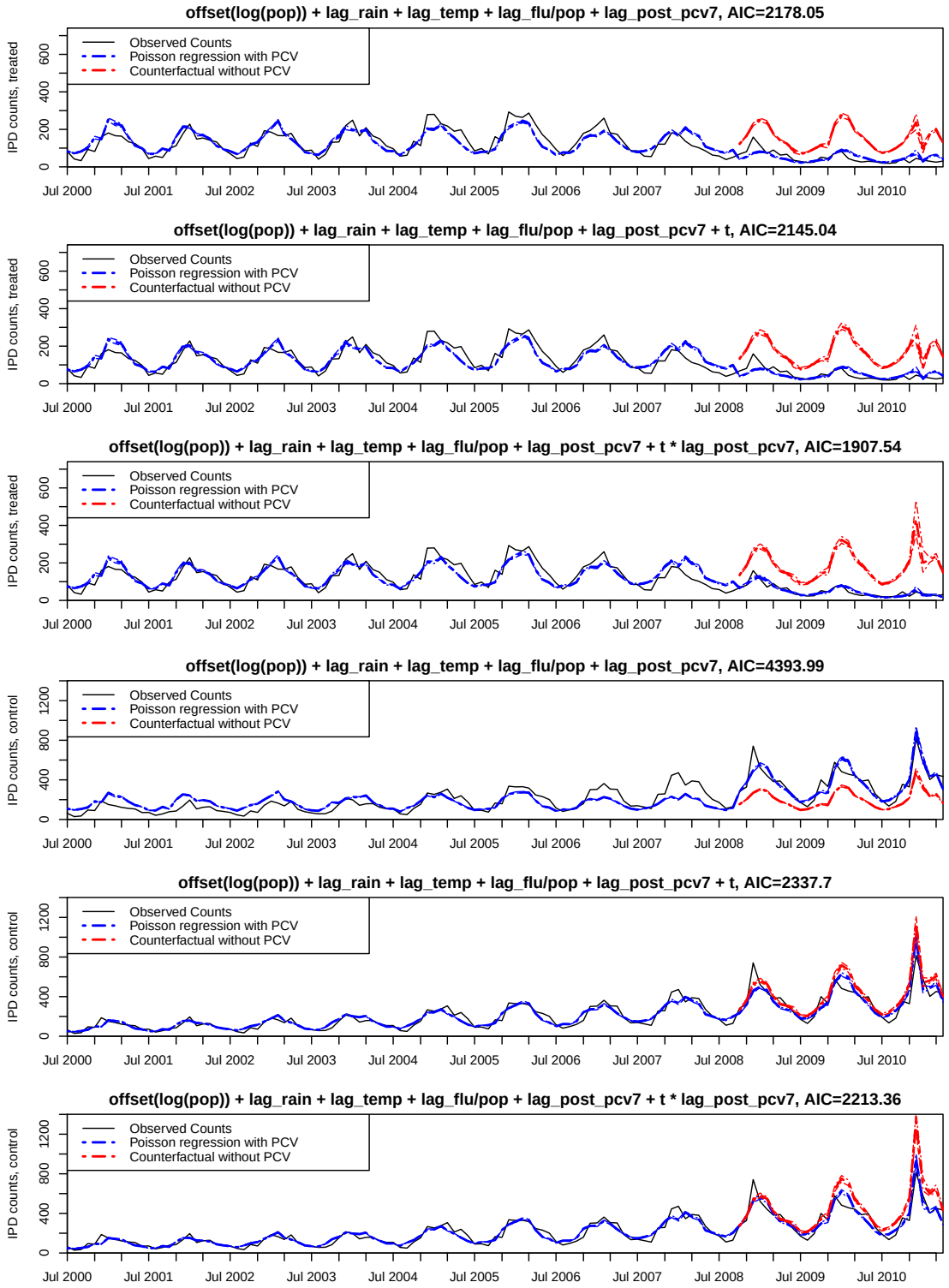


Fig. B.6 Fitted PCV7- and nonPCV7-IPD counts based on three ITS models

.

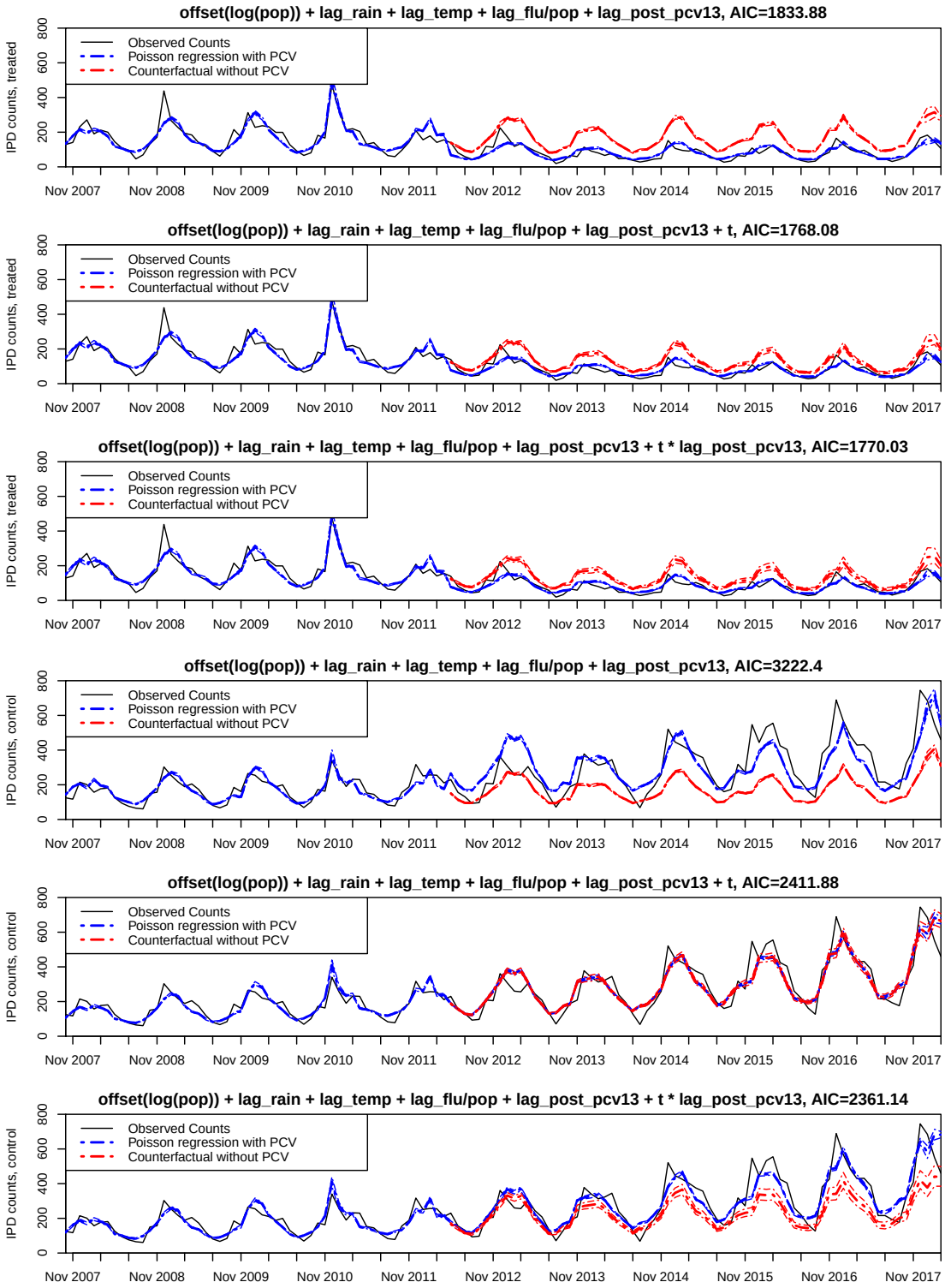


Fig. B.7 Fitted PCV13- and NVT-IPD counts based on three ITS models

B.3 BSTS for IPD by age

B.3.1 Models B and C: impact of PCV7

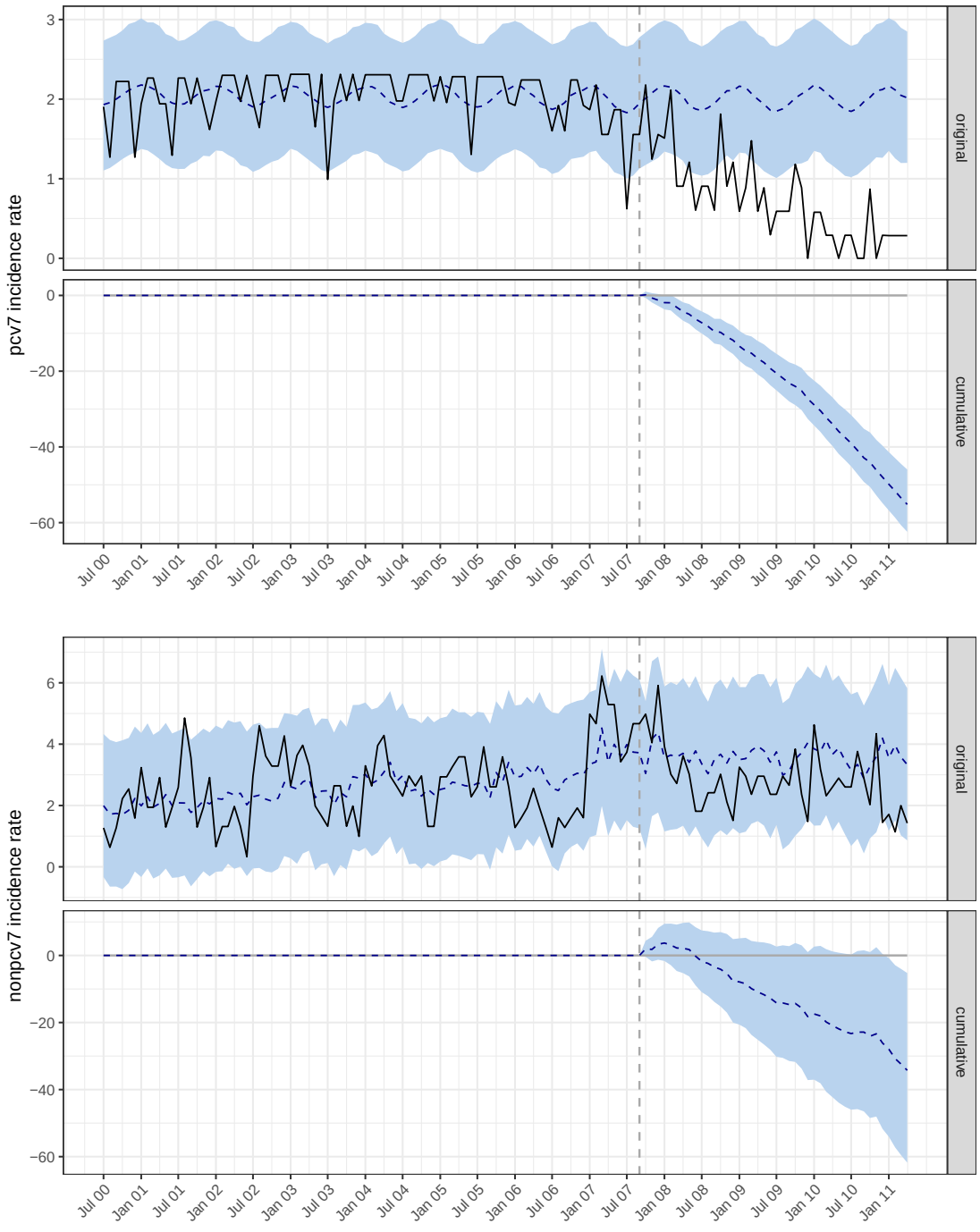


Fig. B.8 Fitted PCV7- and nonPCV7-IPD incidence rates in children younger than 5 .

B.3.2 Models D and E: impact of PCV13

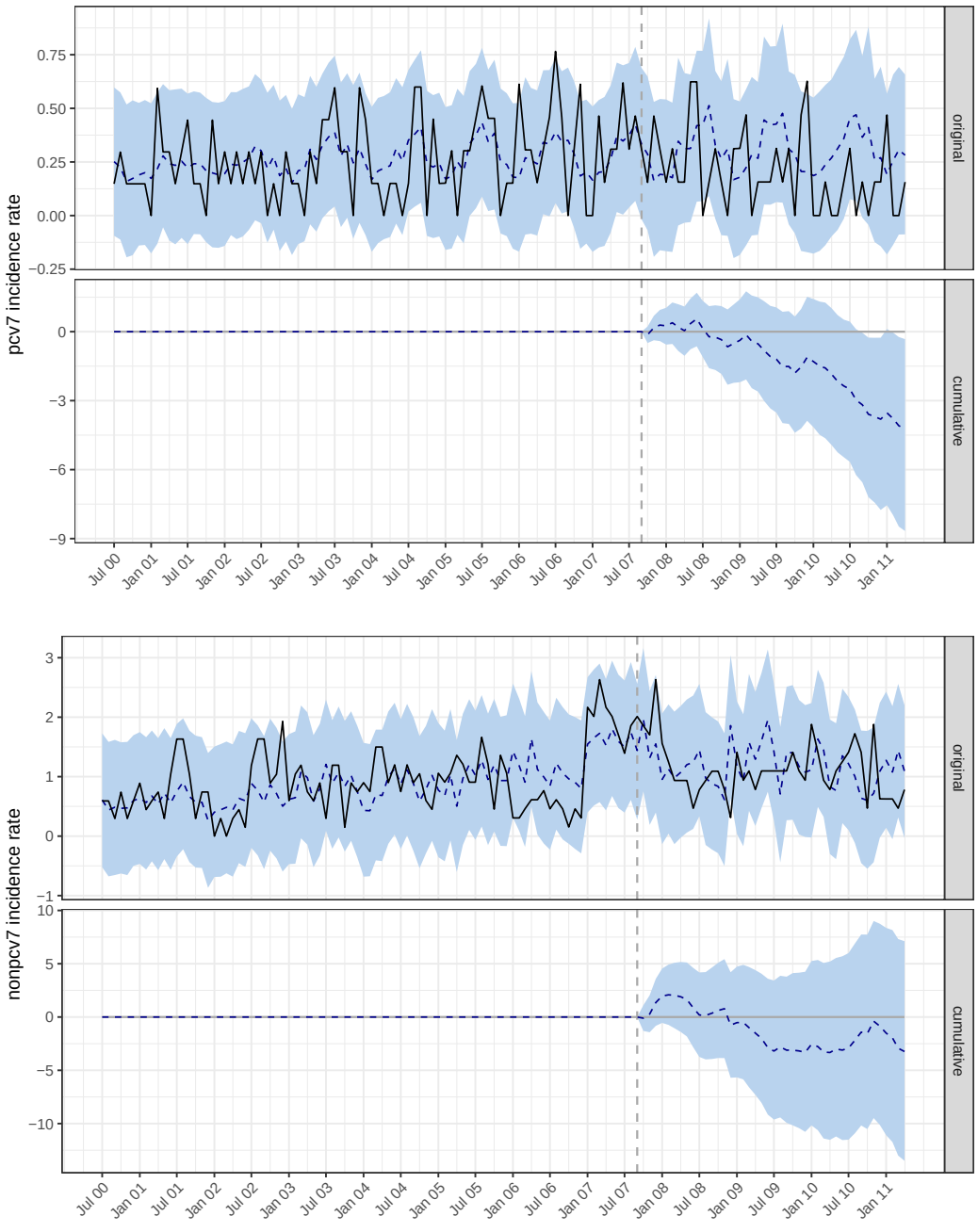


Fig. B.9 Fitted PCV7- and nonPCV7-IPD incidence rates in children of age 5-14

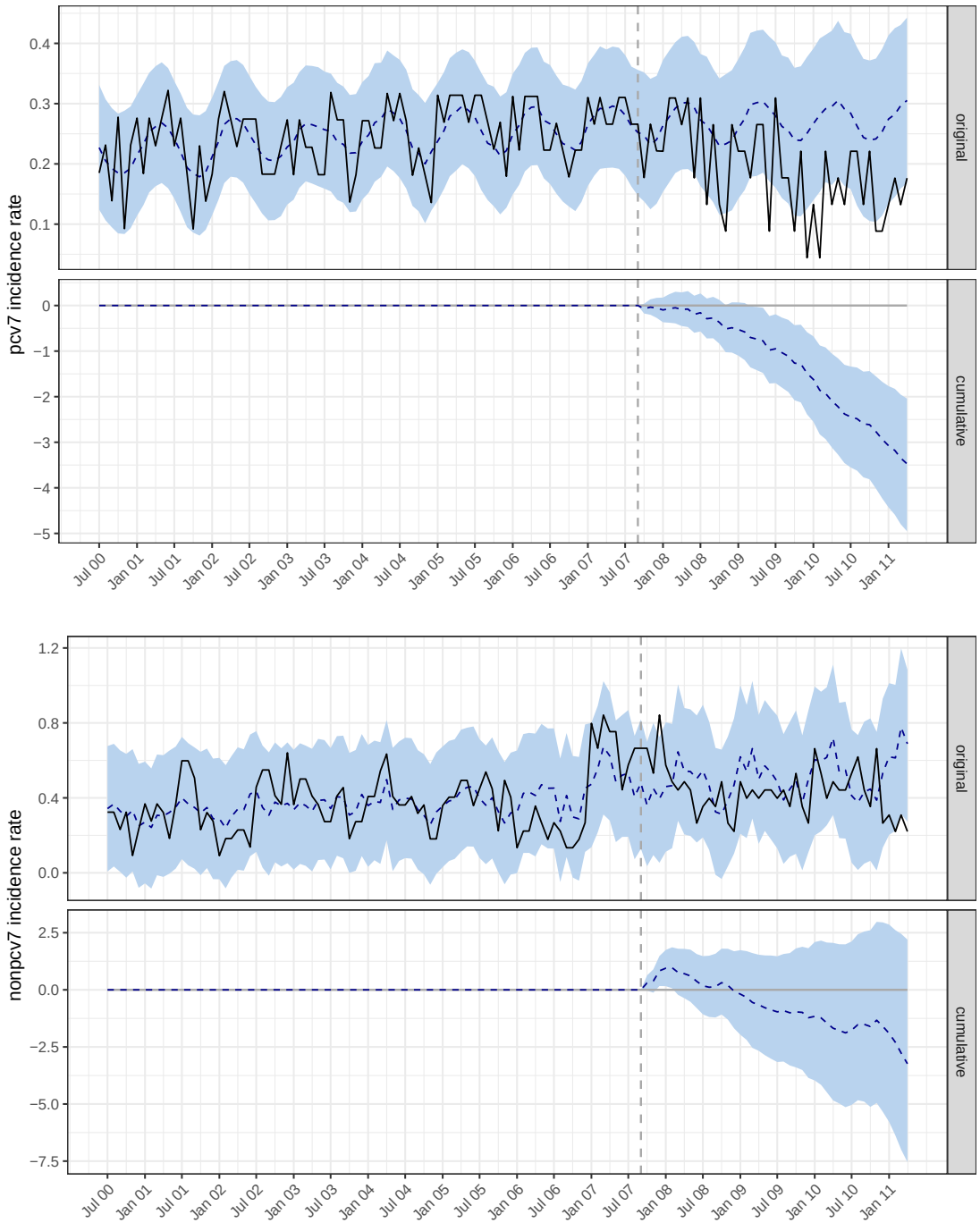


Fig. B.10 Fitted PCV7- and nonPCV7-IPD incidence rates in adults aged 15-44

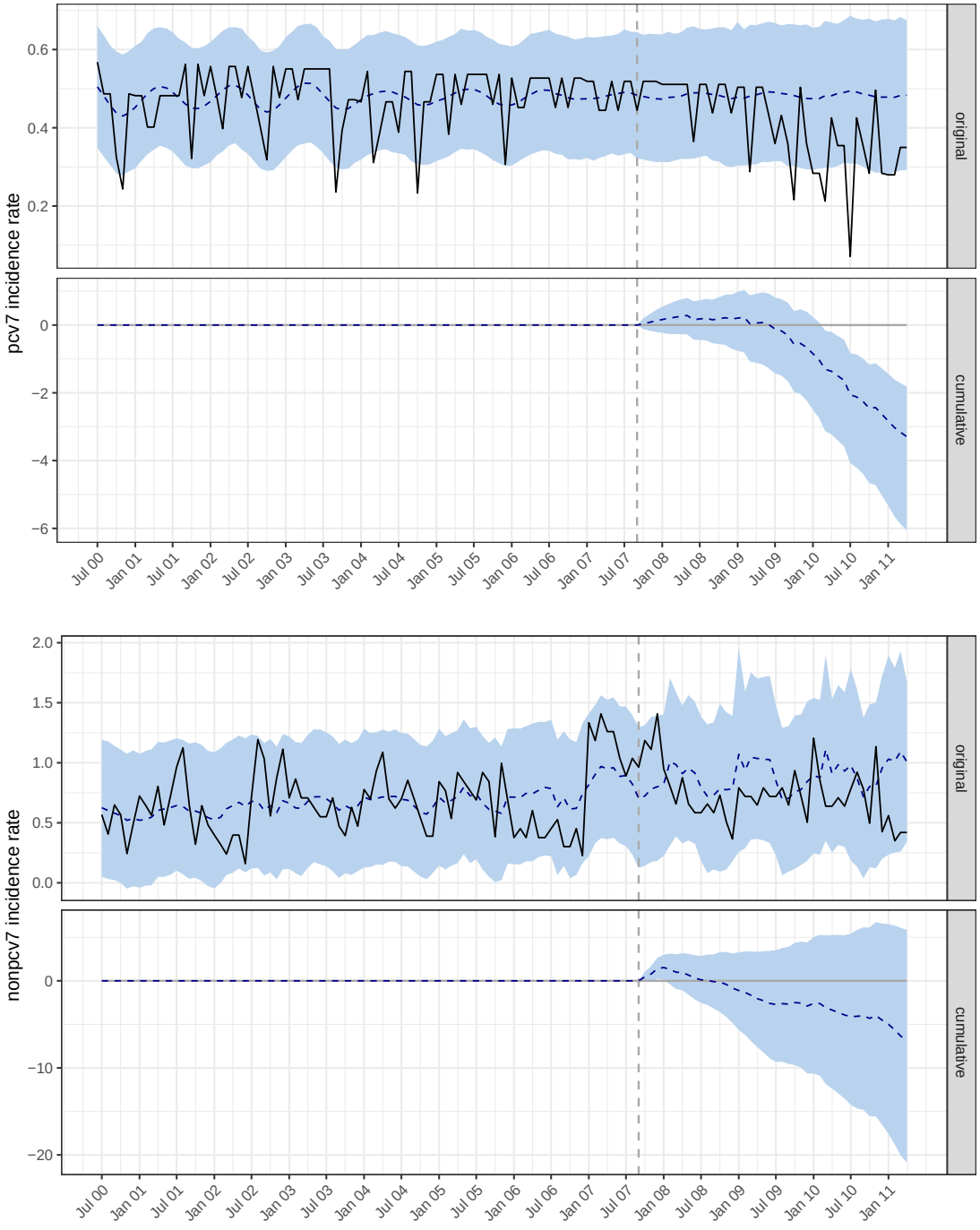


Fig. B.11 Fitted PCV7- and nonPCV7-IPD incidence rates in adults aged 45-64

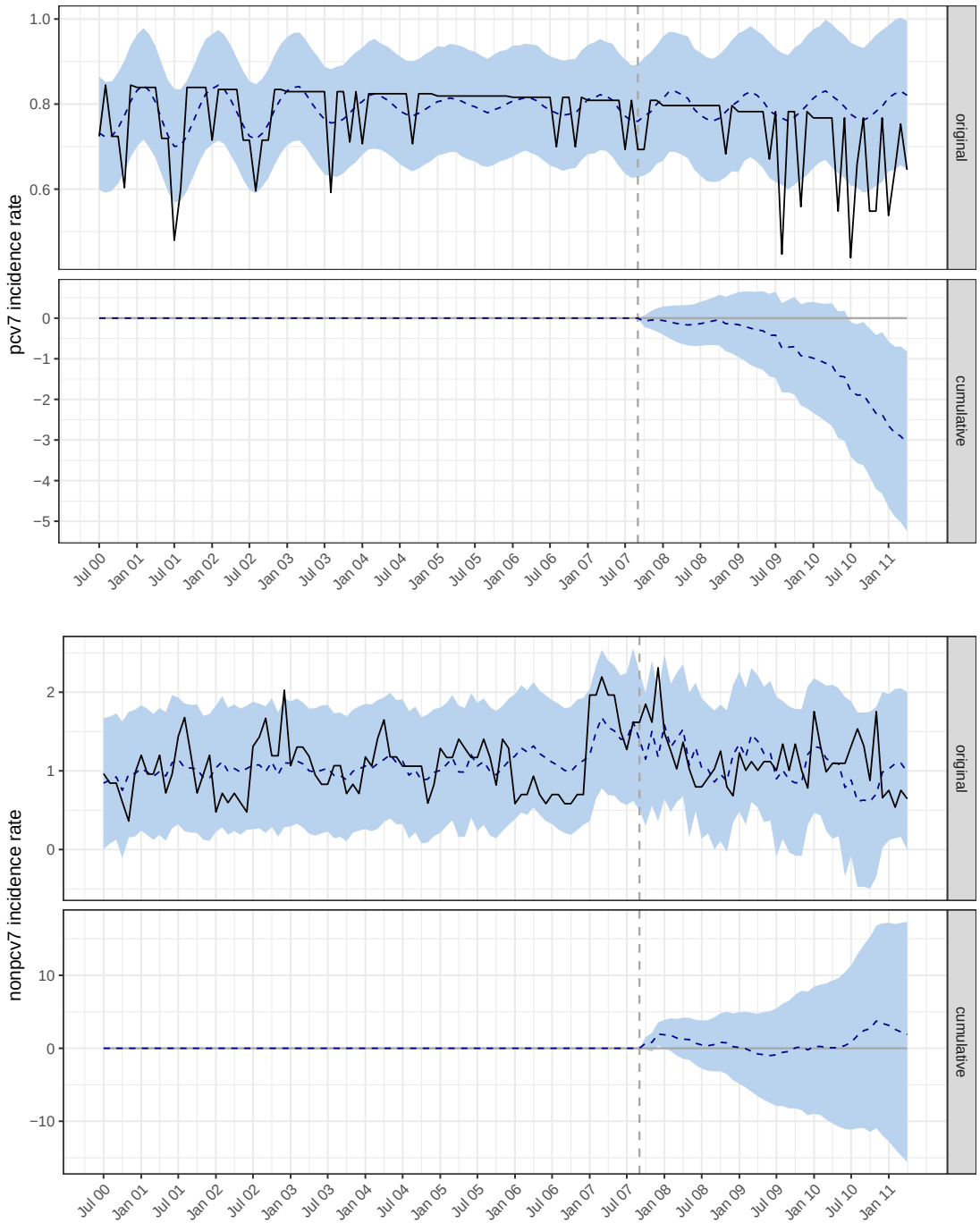


Fig. B.12 Fitted PCV7- and nonPCV7-IPD incidence rates in adults aged 65+

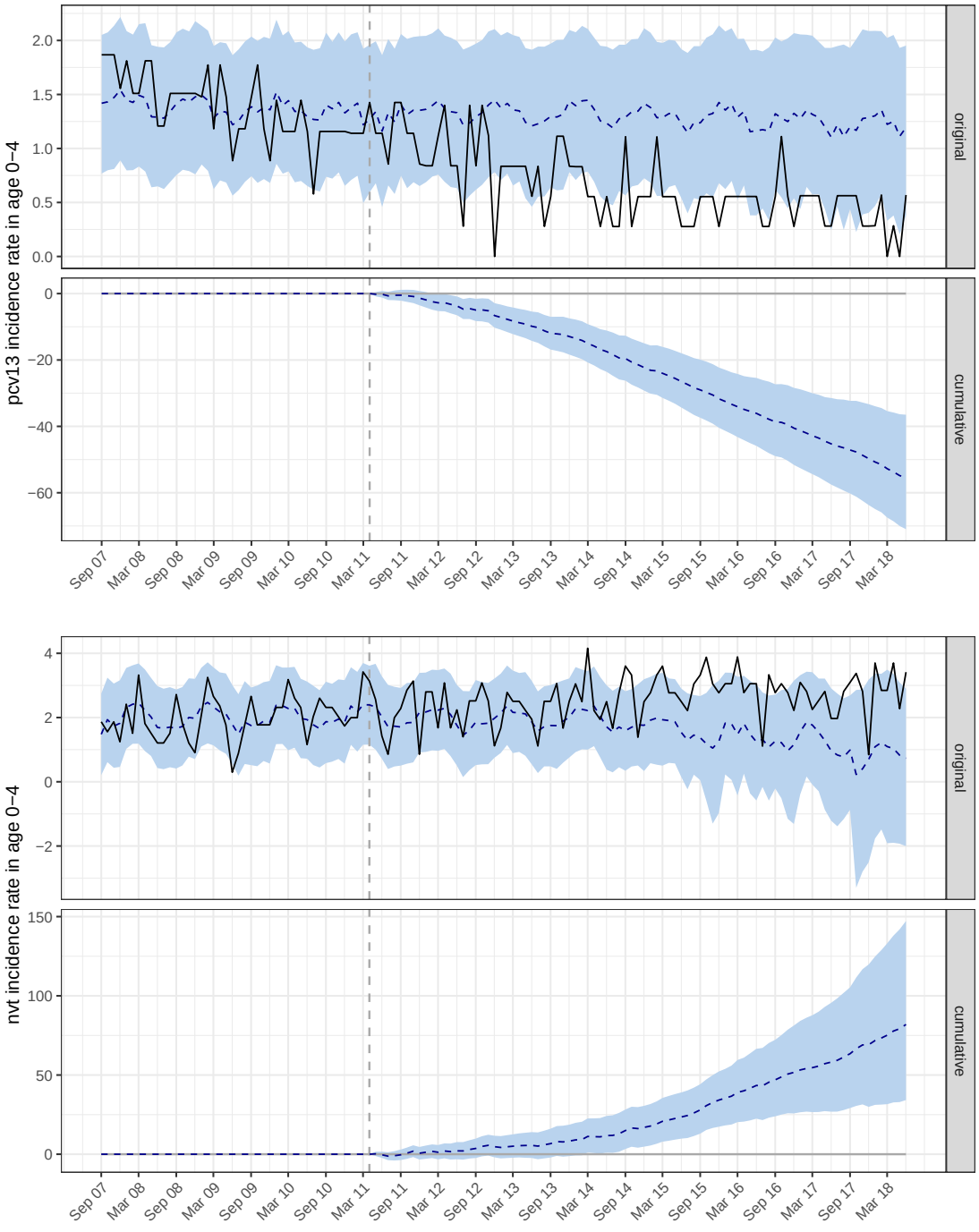


Fig. B.13 Fitted PCV13- and NVT-IPD incidence rates in children younger than 5

.

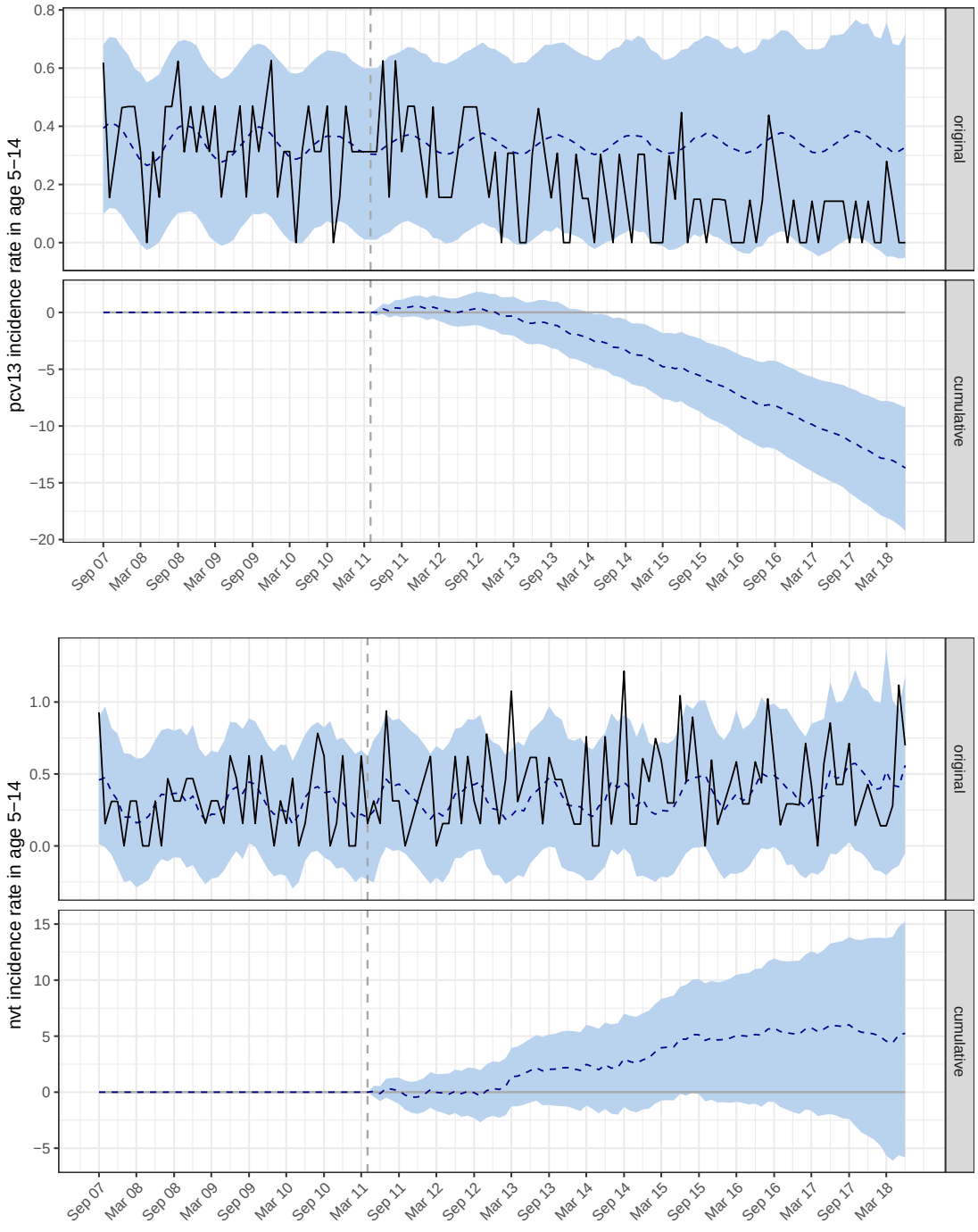


Fig. B.14 Fitted PCV13- and NVT-IPD incidence rates in children of age 5-14

.

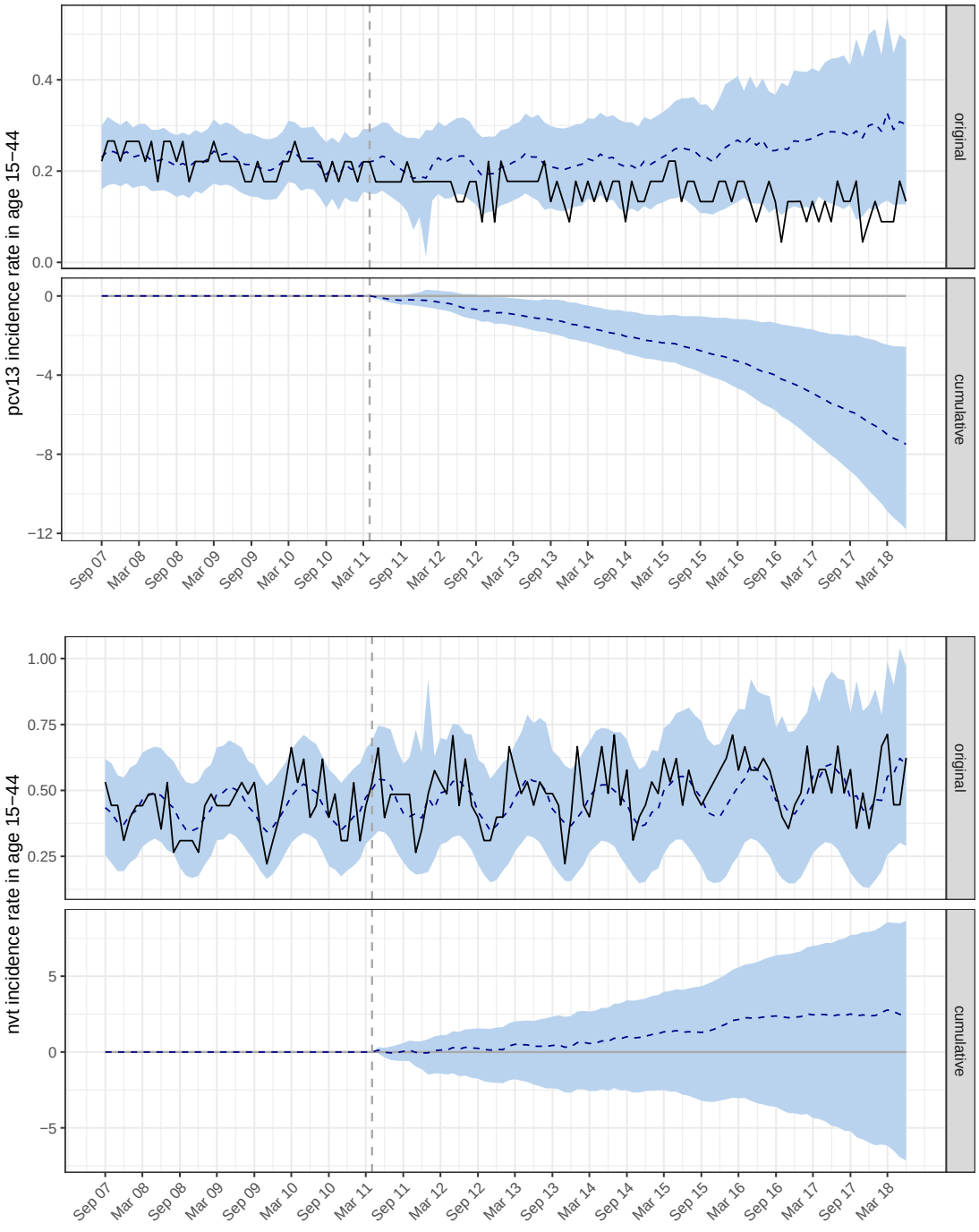


Fig. B.15 Fitted PCV13- and NVT-IPD incidence rates in adults aged 15-44

.

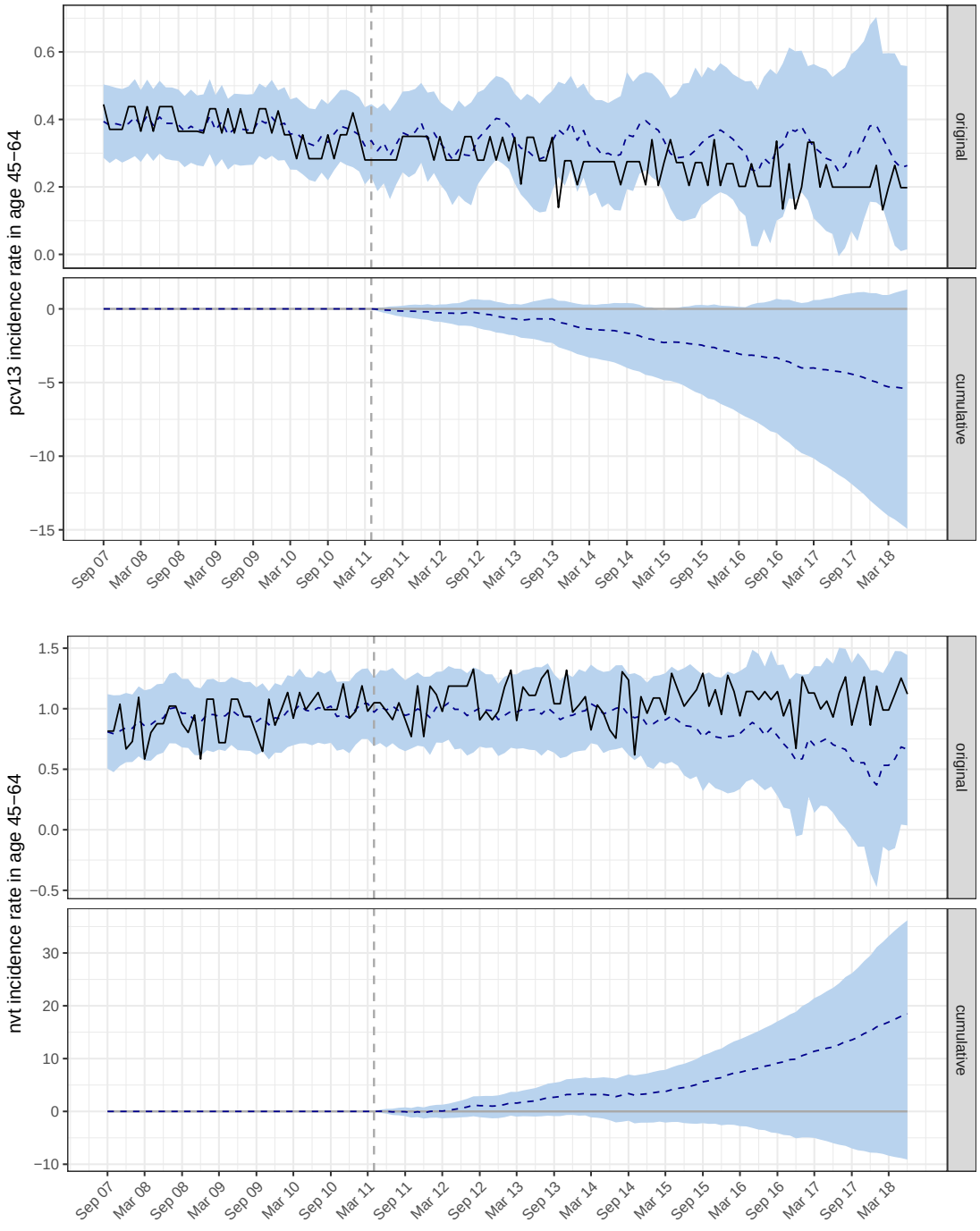


Fig. B.16 Fitted PCV13- and NVT-IPD incidence rates in adults aged 45-64

.

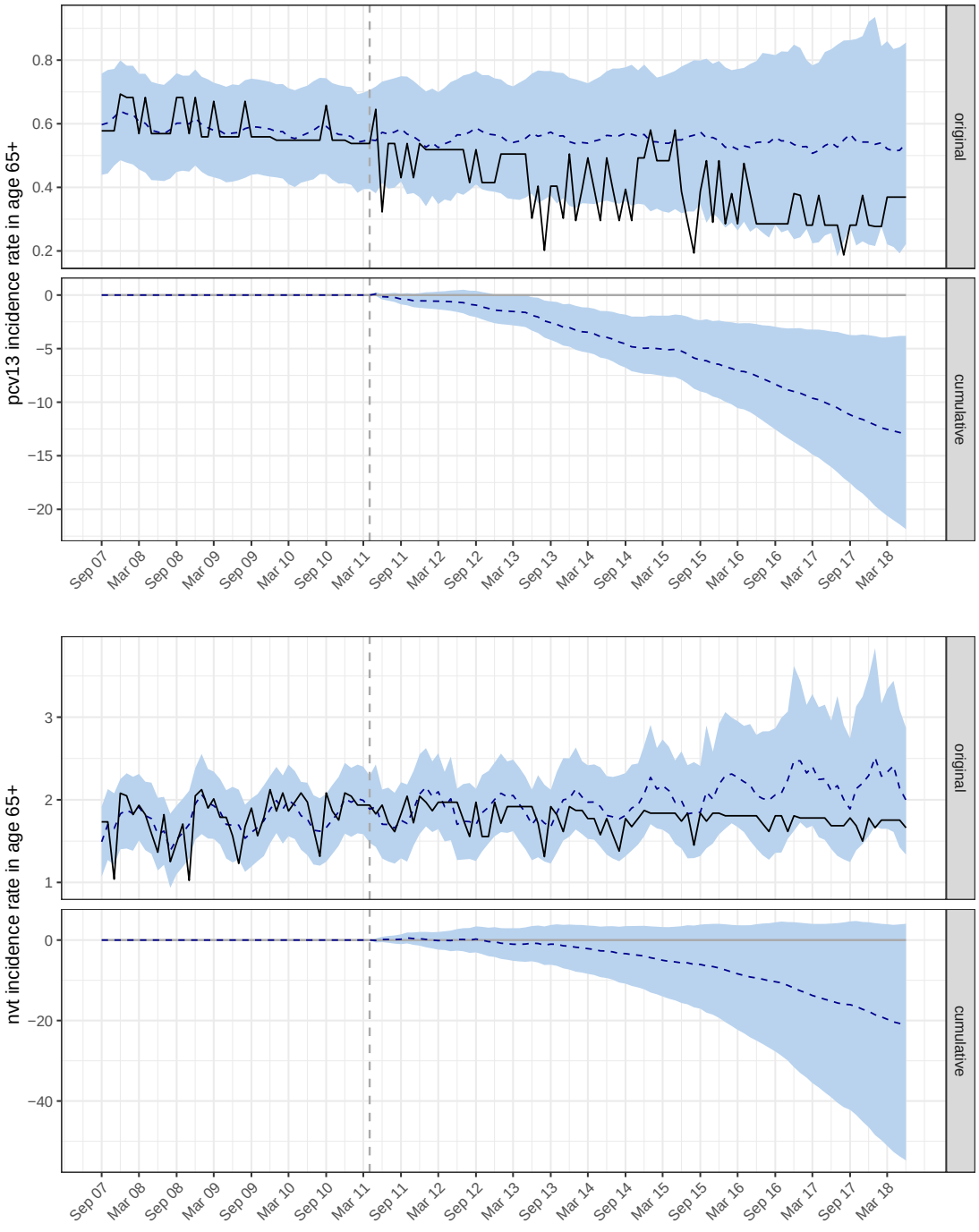


Fig. B.17 Fitted PCV13- and NVT-IPD incidence rates in adults aged 65+

.

Appendix C

Supplementary information to chapter 7

C.1 Singular value decomposition (SVD)

Spectral decomposition, or eigendecomposition, is the expansion of the original data in a coordinate system where the covariate matrix is diagonal. Given a square matrix A, this process leads to its factorization in a simpler form, in terms of its eigenvalues and eigenvectors. Its eigenvalues $\lambda^{(A)}$ and its eigenvectors $\mathbf{x}^{(A)}$ are the set of values for which

$$\mathbf{A}\mathbf{x}^{(A)} = \lambda^{(A)}\mathbf{x}^{(A)} \tag{C.1}$$

or equivalently,

$$\left(\mathbf{A} - \lambda^{(A)}\mathbf{I}\right)\mathbf{x}^{(A)} = 0 \tag{C.2}$$

Since $\mathbf{x}^{(A)}$ must be non-zero, the matrix $\mathbf{A} - \lambda^{(A)}\mathbf{I}$ must have zero determinant. Thus, the eigenvalues $\lambda^{(A)}$ are determined first by solving the characteristic equation $|\mathbf{A} - \lambda^{(A)}\mathbf{I}| = 0$, i.e. by finding a unique set of values $\lambda^{(A)}$ such that the determinant of $\mathbf{A} - \lambda^{(A)}\mathbf{I}$ is equal to zero. The resulting set of eigenvalues is also called spectrum of $\mathbf{A}$.

The set of equations in C.2 is then solved for each eigenvalue: for each value $\lambda_i^{(A)}$ one set of eigenvectors $\mathbf{x}_i^{(A)}$, of dimension n, is obtained. Given n distinct eigenvalues, n sets of linearly independent eigenvectors $\mathbf{x}_i^{(A)}$ are obtained.

This process is also called diagonalisation of $\mathbf{A}$, as finding eigenvalues and eigenvectors allows writing $\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}^{-1}$, where $\mathbf{P}$ is an orthogonal matrix whose columns vector are the set of orthonormal eigenvectors of $\mathbf{A}$, while the diagonal entries of Λ are the corresponding

eigenvalues of **A**. It results that:

$$\mathbf{A} = \lambda_1 p_1 p_1^T + \cdots + \lambda_n p_n p_n^T \tag{C.3}$$

Given a matrix **B**, of dimension *n*x*p*, a generalisation of spectral decomposition is necessary: *singular value decomposition* consists of a two-bases diagonalisation. Eigenvalues and eigenvectors for $\mathbf{BB^T}$ and $\mathbf{B^TB}$ are computed, and this leads to the factorisation

$$\mathbf{B} = \mathbf{U\Sigma V^{-1}} \tag{C.4}$$

where **U** is a *n*x*n* matrix whose columns are the orthonormal eigenvectors of $\mathbf{BB^T}$, also called left singular vectors; Σ is a diagonal matrix *n*x*p* whose diagonal elements are square roots of the eigenvalues of $\mathbf{BB^T}$, also called *singular values*; finally, **V** is a *p*x*p* matrix whose columns are the orthonormal eigenvectors of $\mathbf{B^TB}$, also called right singular vectors. Hence **U** and **V** are orthogonal bases.

Singular values $\sigma_i = \sqrt{\lambda_i}$ are identified in descending order. Generally, $r<p$ of them are positive, where r is the rank of the matrix **B**. It results that $\mathbf{A}v_i = \sigma_i u_i$ for $i = 1, \ldots, r$, or equivalently

$$\mathbf{A} = \sqrt{\lambda_1} u_1 v_1^T + \cdots + \sqrt{\lambda_r} u_r v_r^T \tag{C.5}$$

Hence, the rank r matrix **B** is separated into r rank 1 matrices, called elementary matrices, ordered from the largest to the smallest. The triple $\sqrt{(\lambda_i)}, u_i, v_i$ is also known as ith eigentriple.

C.2 Supplementary results

C.2.1 SSA IPD

When aiming to extract a seasonality component from observed time series, SSA helps to identify eigenvectors that bear correspondence with a harmonic wave. This is easily identifiable from pairs of components which have the form of sine/cosine sequences with the same frequency from the eigenvector plot. In our case, as shown in the top panel of Figure C.1, components 2 and 3 feature four oscillations over 208 weeks, i.e. one oscillation every 52 weeks. Each panel also indicates a percentage representing the portion of total

variance in the observed time series explained by the particular eigenvector: in the case of IPD, seasonality explains a total of 10.5% of the variability.

Visual analysis of the pairwise scatterplots of the singular vectors also allows to visually identify those eigentriples that corresponds to the harmonic components of the series: a pair of sine/cosine waves with equal frequencies, amplitudes, and phases create a scatterplot with points lying on a circle. The purer the information on these harmonic waves, the easier it is to identify their frequency based on the number of vertices of a regular polygon. In our example, eigenvectors 2-3 picture a well-defined circle as they describe yearly periodicity, as shown in the bottom panel of Figure C.1.

C.2.2 SSA flu

In the top panel of Figure C.3, eigenvectors 2 and 3 hint at seasonal harmonics, with four oscillations over 208 weeks as for IPD, jointly explaining 22.7% of the variability. This is confirmed by the plot for the corresponding eigenvector pair, in the bottom panel of Figure C.3, which pictures a circle even though not as well defined as in the case of IPD.

C.2.3 Embedding and CCM on original rates

We show here for comparison the choice of embedding parameters if we had computed AMI and FNN for the original rates instead of the extracted signals: in both cases, we would have required higher embedding dimensions, specifically 4 for IPD and 5 for influenza.

The resulting CCM predictive skills are summarised in Figure C.7

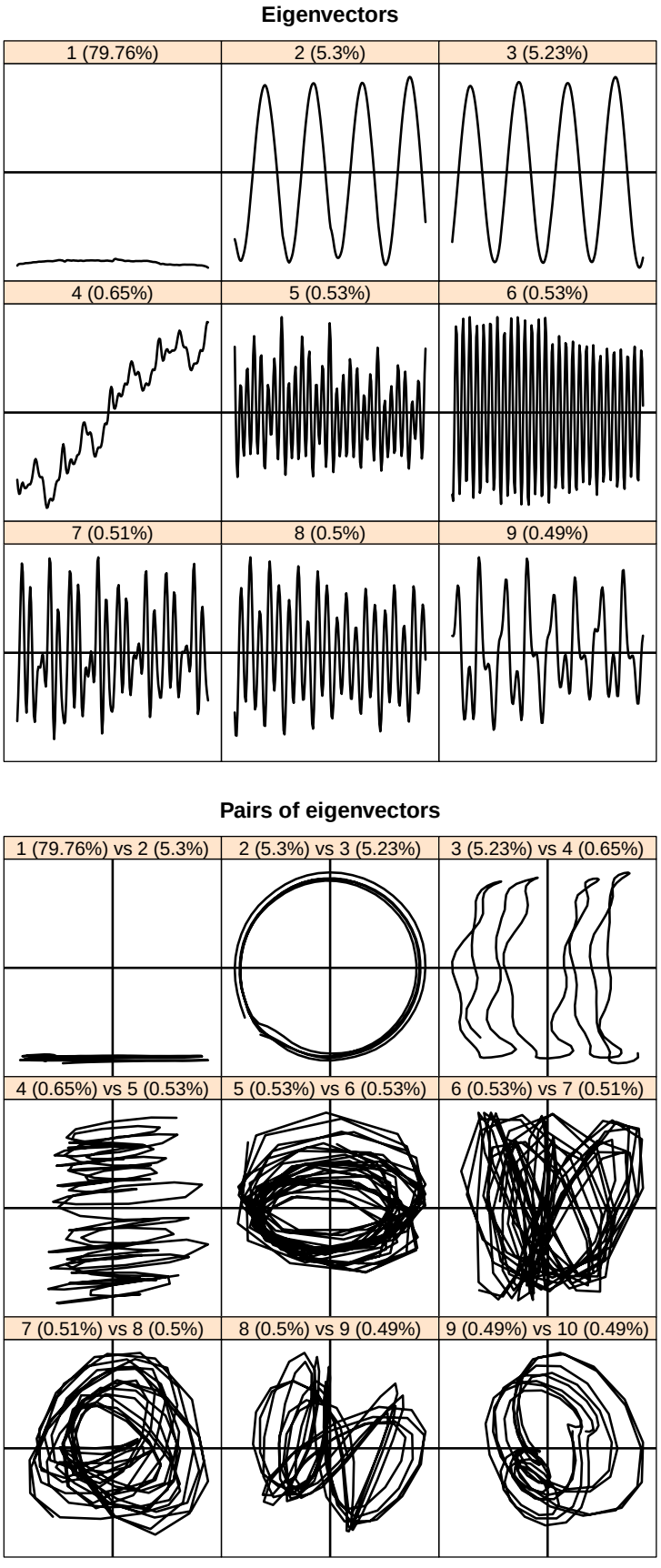


Fig. C.1 Eigenvectors for SSA of IPD time series, individually and in pairs

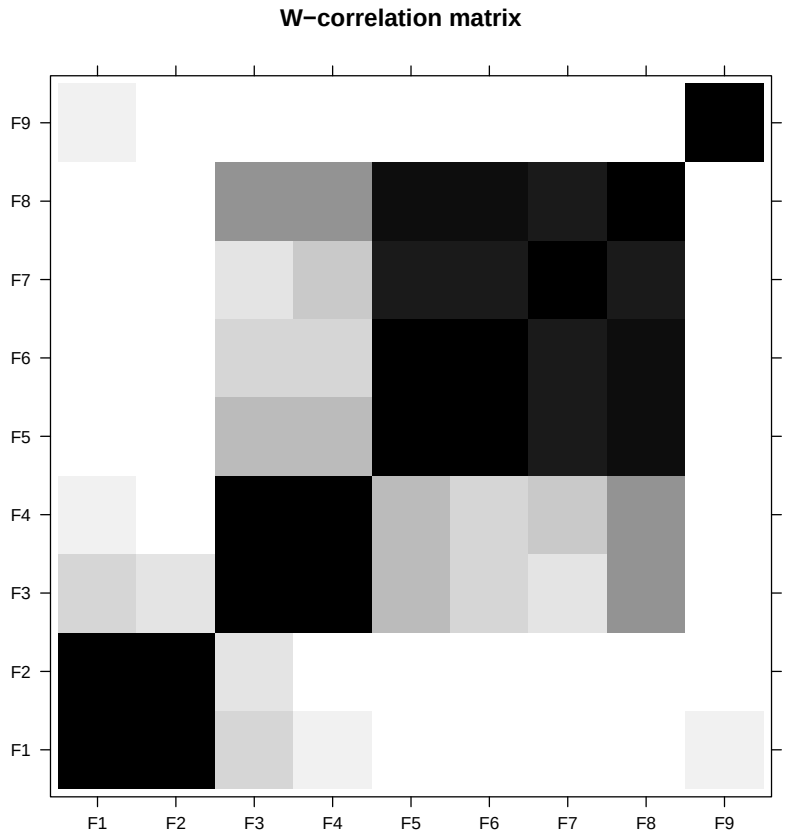


Fig. C.2 W-correlation matrix on the IPD residuals shows no separability for other components

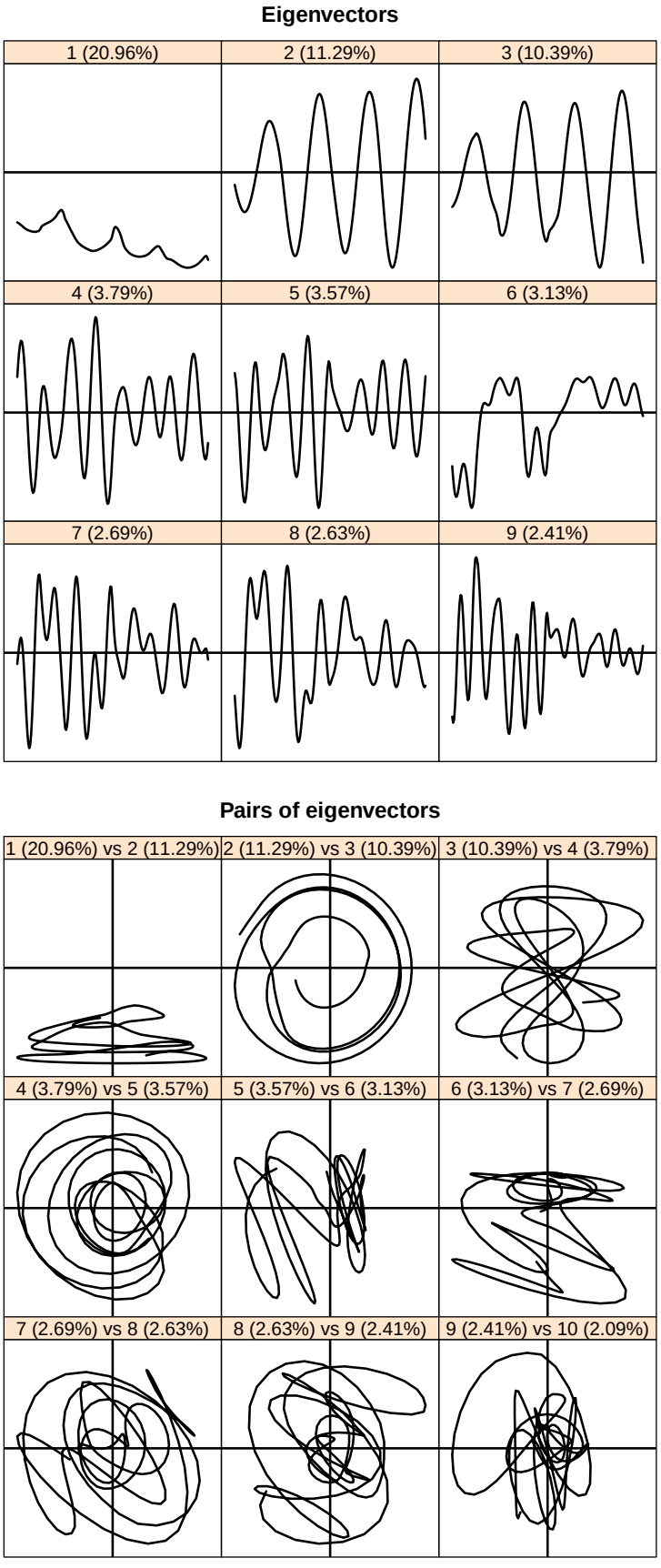


Fig. C.3 Eigenvectors for SSA of Flu time series, individually and in pairs

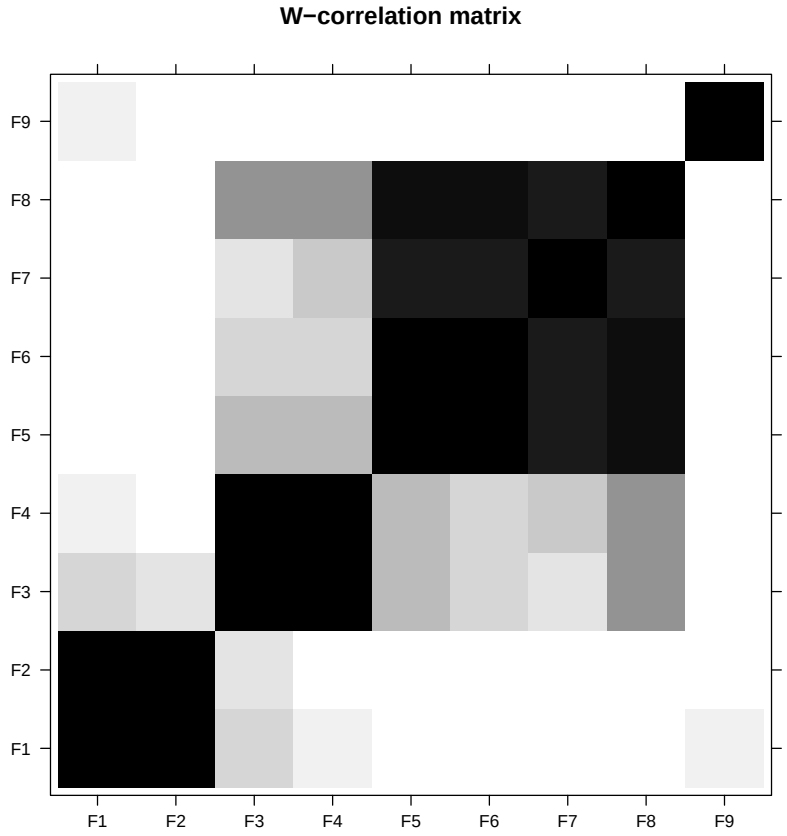


Fig. C.4 W-correlation matrix on the IPD residuals shows no separability for other components

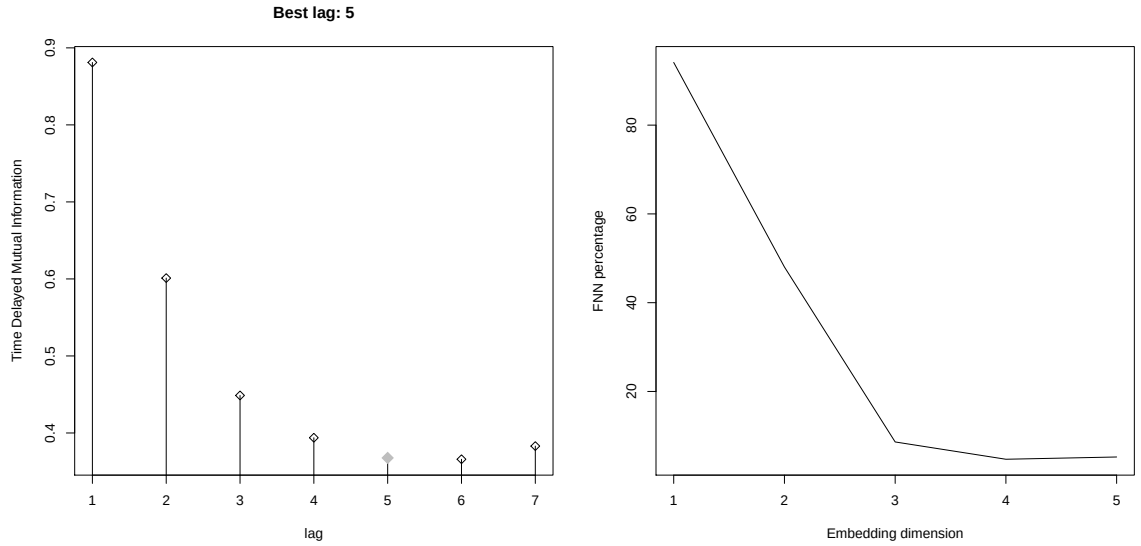


Fig. C.5 Selection of time delay τ and embedding dimension E for the original IPD rates

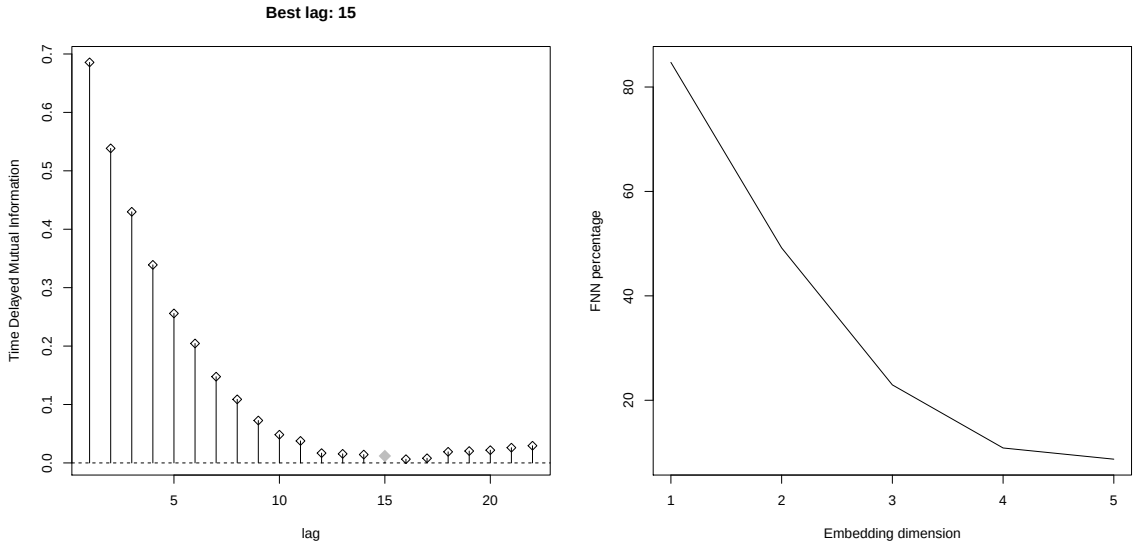


Fig. C.6 Selection of time delay τ and embedding dimension E for the original flu rates

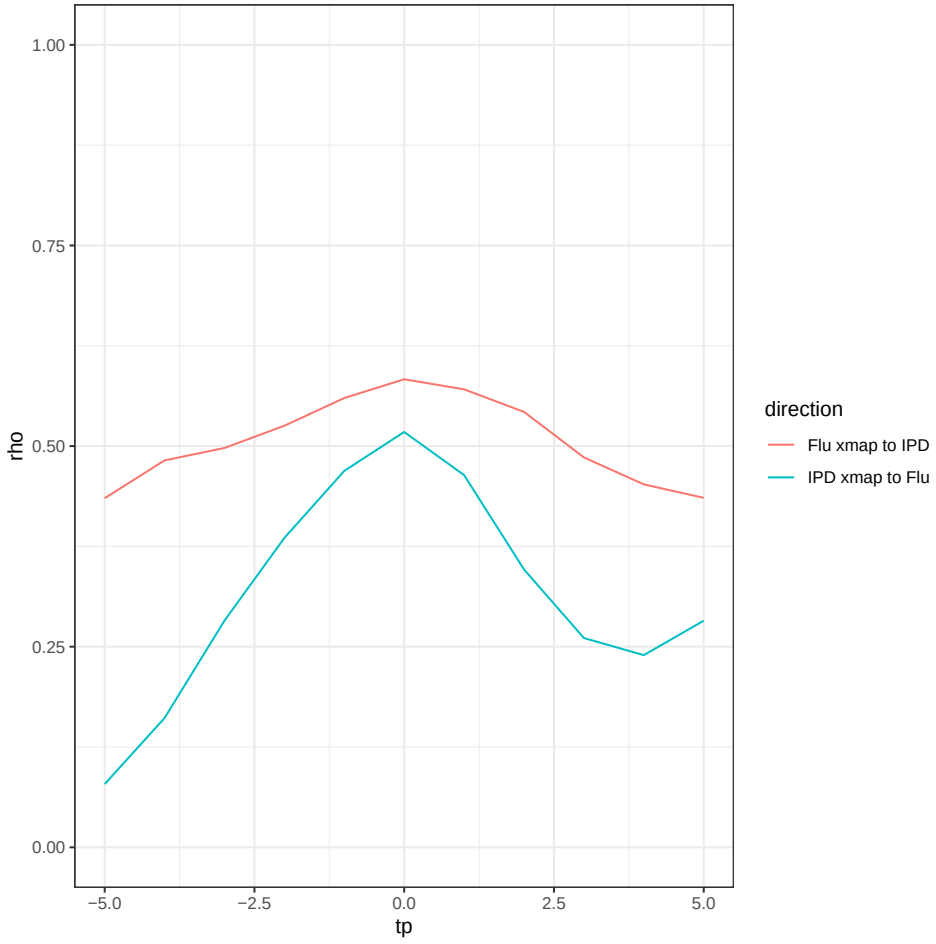


Fig. C.7 CCM predictive skills when applied on the original time series